PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

| (51) International Patent Classification [6] : C12Q 1/68, C07H 21/00	A1	(11) International Publication Number: WO 96/12039
		(43) International Publication Date: 25 April 1996 (25.04.96)

(21) International Application Number: PCT/US95/12678

(22) International Filing Date: 12 October 1995 (12.10.95)

(30) Priority Data:
| 08/322,348 | 13 October 1994 (13.10.94) | US |
| 08/359,295 | 19 December 1994 (19.12.94) | US |

(71) Applicant: LYNX THERAPEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US).

(72) Inventor: BRENNER, Sydney; University of Cambridge, School of Clinical Medicine, Level 5, Addenbrooke's Hospital, Hills Road, Cambridge CB 2QQ (GB).

(74) Agent: MACEVICZ, Stephen, C.; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).

(81) Designated States: AU, CA, CZ, FI, HU, JP, KR, NO, SG, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

(54) Title: MASSIVELY PARALLEL SEQUENCING OF SORTED POLYNUCLEOTIDES

(57) Abstract

The invention provides a method and materials for sorting polynucleotides with oligonucleotide tags. Oligonucleotide tags of the invention are capable of hybridizing to complementary oligomeric compounds consisting of subunits having enhanced binding strength and specificity as compared to natural oligonucleotides. Such complementary oligomeric compounds are referred to herein as "tag complements". Subunits of tag complements may consist of monomers of non-natural nucleotide analogs, referred to herein as "antisense monomers" or they may comprise oligomers having lengths in the range of 3 to 6 nucleotides or analogs thereof, including antisense monomers, the oligomers being selected from a minimally cross-hybridizing set. In such a set, a duplex made up of an oligomer of the set and the complement of any other oligomer of the set contains at least two mismatches. Preferred antisense monomers include peptide nucleic acid monomers and nucleoside phosphoramidates having a 3'-NHP(O)(O-)O-5' linkage with its adjacent nucleoside. An important aspect of the invention is the use of the oligonucleotide tags for sorting polynucleotides by specifically hybridizing tags attached to the polynucleotides to their complements on solid phase supports. This embodiment provides a readily automated system for manipulating and sorting polynucleotides, particularly useful in large-scale parallel operations, such as large-scale DNA sequencing, mRNA fingerprinting, or the like, wherein many target polynucleotides or many segments of a single target polynucleotide are sequenced simultaneously.

PROBE

AUGMENTED PROBE
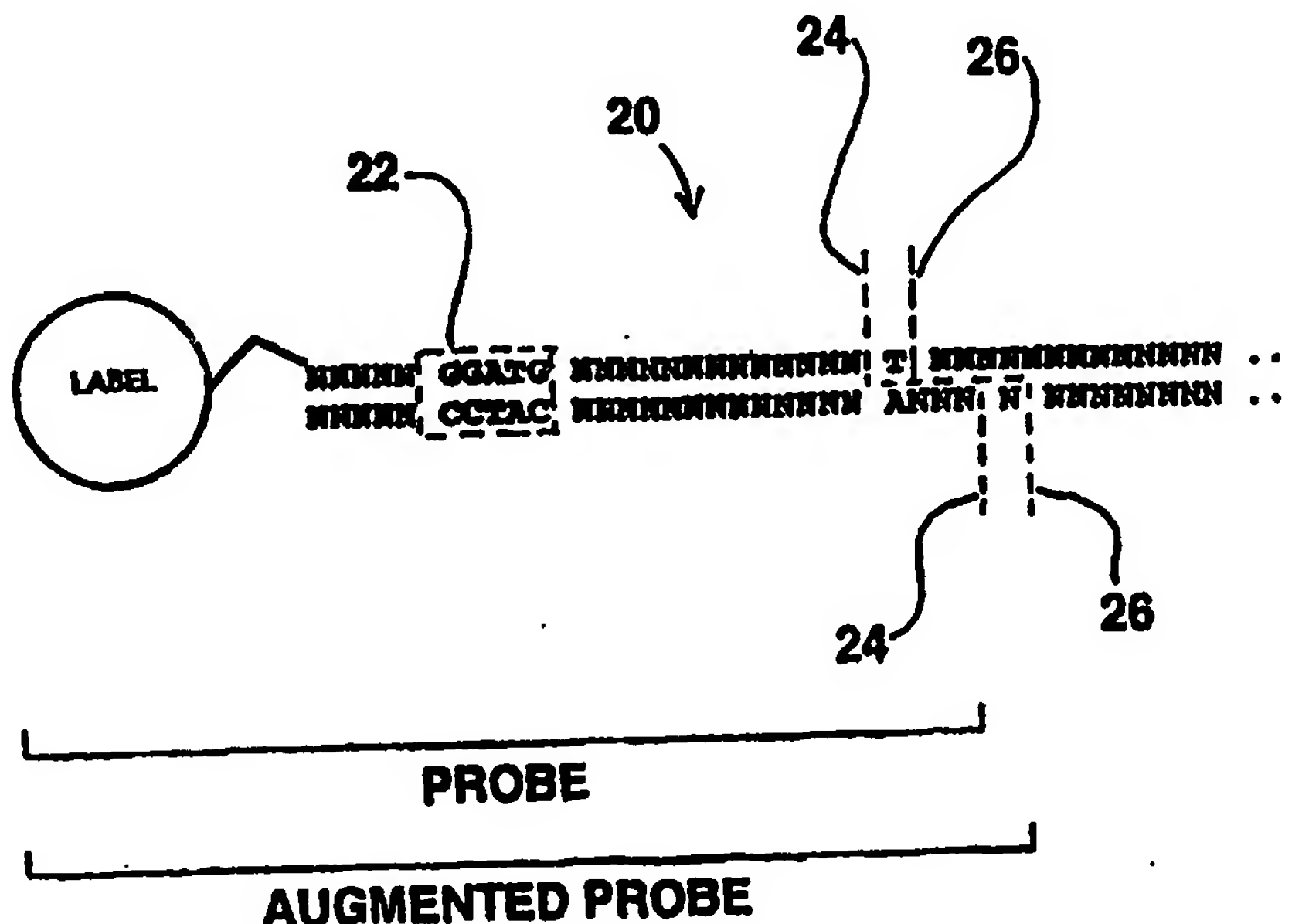
FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic	SD	Sudan
CG	Congo		of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

MASSIVELY PARALLEL SEQUENCING OF SORTED POLYNUCLEOTIDES

Field of the Invention

The invention relates generally to methods for sequencing polynucleotides, and more particularly, to a method of sorting and sequencing many polynucleotides simultaneously.

BACKGROUND

Presently there are two basic approaches to DNA sequence determination: the chain termination method, e.g. Sanger et al, Proc. Natl. Acad. Sci., 74: 5463-5467 (1977); and the chemical degradation method, e.g. Maxam et al, Proc. Natl. Acad. Sci., 74: 560-564 (1977). The chain termination method has been improved in many ways since its invention, and serves as the basis for all currently available automated DNA sequencing machines, e.g. Sanger et al, J. Mol. Biol., 143: 161-178 (1980); Schreier et al, J. Mol. Biol., 129: 169-172 (1979); Smith et al, Nature, 321: 674-679 (1987); Prober et al, Science, 238: 336-341 (1987); Hunkapiller et al, Science, 254: 59-67 (1991); Bevan et al, PCR Methods and Applications, 1: 222-228 (1992). Moreover, further improvements are easily envisioned that should greatly enhance the throughput and efficiency of the approach, e.g. Huang et al, Anal. Chem., 64: 2149-2154 (1992)(capillary arrays); Best et al, Anal. Chem., 66: 4063-4067 (1994)(non-cross-linked polymeric separation media for capillaries); better dye sets; and the like.

Nonetheless, even with such reasonably envisioned improvements, these approaches still have several inherent technical problems that make them both expensive and time consuming, particularly when applied to large-scale sequencing projects. Such problems include i) the gel electrophoretic separation step which is labor intensive, is difficult to automate, and which introduces an extra degree of variability in the analysis of data, e.g. band broadening due to temperature effects, compressions due to secondary structure in the DNA sequencing fragments, inhomogeneities in the separation gel, and the like; ii) nucleic acid polymerases whose properties, such as processivity, fidelity, rate of polymerization, rate of incorporation of chain terminators, and the like, are often sequence dependent; iii) detection and analysis of DNA sequencing fragments which are typically present in fmol quantities in spatially overlapping bands in a gel; iv) lower signals because the labeling moiety is distributed over the many hundred spatially separated bands rather

than being concentrated in a single homogeneous phase, v) in the case of single-lane fluorescence detection, the availability of dyes with suitable emission and absorption properties, quantum yield, and spectral resolvability; and vi) the need for a separately prepared sequencing template for each sequencing reaction to identify a maximum of

5 about 400-600 bases, e.g. Trainor, Anal. Biochem., 62: 418-426 (1990); Connell et al, Biotechniques, 5: 342-348 (1987); Karger et al, Nucleic Acids Research, 19: 4955-4962 (1991); Fung et al, U.S. patent 4,855,225; Nishikawa et al, Electrophoresis, 12: 623-631 (1991); and Hunkapiller et al (cited above).

The need to prepare separate sequencing templates is especially onerous in

10 large-scale sequencing projects, e.g. Hunkapiller et al (cited above)(94.4 kilobase target--2399 templates); and Alderton et al, Anal. Biochem., 201: 166-169 (1992)(230 kilobase target--13,000 templates). Attempts to automate template preparation have proved difficult, especially when coupled with current sequencing methodologies, e.g. Church et al, Science, 240: 185-188 (1988); Beck et al, Anal.

15 Biochem. 212: 498-505 (1993); Wilson et al, Biotechniques, 6: 776-787 (1988); and the like.

In view of the above, a major advance in sequencing technology would take place if there were means available for overcoming the template-preparation bottleneck. In particular, the ability to prepare many thousands of templates

20 simultaneously without individual template selection and handling would lead to significant increases in sequencing throughput and significant lowering of sequencing costs.

Summary of the Invention

25 An object of my invention is to provide a method for tagging and sorting many thousands of fragments of a target polynucleotide for simultaneous analysis and/or sequencing.

Another object of my invention is to provide a method, kits, and apparatus for analyzing and/or sequencing many thousands of different polynucleotides

30 simultaneously.

A further object of my invention is to provide a method for greatly reducing the number of separate template preparation steps required in large scale sequencing projects.

Still another object of my invention is to provide a method for applying

35 single-base sequencing methodologies to many different target polynucleotides simultaneously.

Another object of my invention is to provide a rapid and reliable method for sequencing target polynucleotides having a length in the range of a few hundred basepairs to several tens of thousands of basepairs.

My invention achieves these and other objects by providing a method and
5   materials for sorting polynucleotides with oligonucleotide tags. Oligonucleotide tags of the invention are capable of hybridizing to complementary oligomeric compounds consisting of subunits having enhanced binding strength and specificity as compared to natural oligonucleotides. Such complementary oligomeric compounds are referred to herein as "tag complements." Subunits of tag complements may consist of
10   monomers of non-natural nucleotide analogs, referred to herein as "antisense monomers" or they may comprise oligomers having lengths in the range of 3 to 6 nucleotides or analogs thereof, including antisense monomers, the oligomers being selected from a minimally cross-hybridizing set. In such a set, a duplex made up of an oligomer of the set and the complement of any other oligomer of the set contains at
15   least two mismatches. In other words, an oligomer of a minimally cross-hybridizing set at best forms a duplex having at least two mismatches with the complement of any other oligomer of the same set. The number of oligonucleotide tags available in a particular embodiment depends on the number of subunits per tag and on the length of the subunit, when the subunit is an oligomer from a minimally cross-hybridizing
20   set. In the latter case, the number is generally much less than the number of all possible sequences the length of the tag, which for a tag n nucleotides long would be 4^n. Preferred antisense monomers include peptide nucleic acid monomers and nucleoside phosphoramidates having a 3'-NHP(O)(O-)O-5' linkage with its adjacent nucleoside. The latter compounds are referred to herein as 3'N$\rightarrow$5'P
25   phosphoramidates.

In one aspect of my invention, tag complements attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. In this embodiment, tag complements are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location in an array
30   of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of tag complement which has a particular sequence. The population of such beads or regions contains a repertoire of tag complements with distinct sequences, the size of the
35   repertoire depending on the number of subunits per oligonucleotide tag and the length of the subunits employed, where oligomeric subunits are used. Similarly, the polynucleotides to be sorted each comprises an oligonucleotide tag in the repertoire, such that identical polynucleotides have the same tag and different polynucleotides

have different tags. As explained more fully below, this condition is achieved by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. Thus, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide

5   tags with their respective complements, subpopulations of identical polynucleotides are sorted onto particular beads or regions. The subpopulations of polynucleotides can then be manipulated on the solid phase support by micro-biochemical techniques.

An important aspect of my invention is the use of the oligonucleotide tags to sort polynucleotides for parallel sequence determination. Preferably, this aspect of

10 the invention comprises the following steps: (a) generating from a target polynucleotide a plurality of fragments that covers the target polynucleotide; (b) attaching an oligonucleotide tag from a repertoire of tags to each fragment of the plurality (i) such that substantially all the same fragments have the same oligonucleotide tag attached and (ii) such that each oligonucleotide tag from the

15 repertoire comprises a plurality of subunits and each subunit of the plurality consists of a complementary nucleotide of an antisense monomer or an oligonucleotide having a length from three to six nucleotides, the oligonucleotides being selected from a minimally cross-hybridizing set; (c) sorting the fragments by specifically hybridizing the oligonucleotide tags with their respective tag complements; (d) determining the

20 nucleotide sequence of a portion of each of the fragments of the plurality; and (e) determining the nucleotide sequence of the target polynucleotide by collating the sequences of the fragments.

Another important feature of my invention is a method of identifying, or fingerprinting, a population of mRNA or cDNA molecules. Preferably, such a

25 method comprises the following steps: (a) forming a population of cDNA molecules from the population of mRNA molecules, the cDNA molecules being complementary to the mRNA molecules and each cDNA molecule having an oligonucleotide tag attached, (i) such that substantially all of the same cDNA molecules have the same oligonucleotide tag attached and (ii) such that each oligonucleotide tag from the

30 repertoire comprises a plurality of subunits and each subunit of the plurality consists of a complementary nucleotide of an antisense monomer or an oligonucleotide having a length from three to six nucleotides, the oligonucleotides being selected from a minimally cross-hybridizing set; (b) sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective tag complements; (c)

35 determining the nucleotide sequence of a portion of each of the sorted cDNA molecules; and (d) identifying the population of mRNA molecules by the frequency distribution of the portions of sequences of the cDNA molecules. Clearly, this embodiment of the invention may also be used to identify novel polynucleotides in

- 4 -

cDNA libraries. In this latter embodiment, step (d) would be identifying a novel cDNA molecule by comparing the sequences of the portions of the sorted cDNAs with known sequences, e.g. via sequence libraries, such as Genbank.

5   When used in combination with solid phase supports, such as microscopic beads, my invention provides a readily automated system for manipulating and sorting polynucleotides, particularly useful in large-scale parallel operations, such as large-scale DNA sequencing, mRNA fingerprinting, and the like, where many target polynucleotides or many segments of a single target polynucleotide are sequenced and/or analyzed simultaneously.

10

Brief Description of the Drawings

        Figures 1a-1c illustrates structures of labeled probes employed in a preferred method of "single base" sequencing which may be used with the invention.

        Figure 2 illustrates the relative positions of the nuclease recognition site,
15   ligation site, and cleavage site in a ligated complex formed between a target polynucleotide and a probe used in a preferred "single base" sequencing method.

        Figure 3 is a flow chart illustrating a general algorithm for generating minimally cross-hybridizing sets.

        Figure 4 illustrates a scheme for synthesizing oligonucleotide $3'N \rightarrow 5'P$
20   phosphoramidates.

        Figure 5 diagrammatically illustrates an apparatus for carrying out parallel operations, such as polynucleotide sequencing, in accordance with the invention.

Definitions

25   "Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are
30   formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

        The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides,
35   α-anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually

monomers are linked by phosphodiester bonds or analogs thereof to form
oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens
of monomeric units. Whenever an oligonucleotide is represented by a sequence of
letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3'
5   order from left to right and that "A" denotes deoxyadenosine, "C" denotes
deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless
otherwise noted. Analogs of phosphodiester linkages include phosphorothioate,
phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. It is clear to
those skilled in the art when oligonucleotides having natural or non-natural
10   nucleotides may be employed, e.g. where processing by enzymes is called for, usually
oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or
oligonucleotide strands making up the duplex form a double stranded structure with
one other such that every nucleotide in each strand undergoes Watson-Crick
15   basepairing with a nucleotide in the other strand. The term also comprehends the
pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine
bases, and the like, that may be employed. In reference to a triplex, the term means
that the triplex consists of a perfectly matched duplex and a third strand in which
every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a
20   basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex
between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the
duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse
Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-
25   deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA
Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to
nucleosides includes synthetic nucleosides having modified base moieties and/or
modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley,
New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or
30   the like, with the only proviso that they are capable of specific hybridization. Such
analogs include synthetic nucleosides designed to enhance binding properties, reduce
degeneracy, increase specificity, and the like.

"Stable" in reference to the formation of a covalent linkage and/or non-
covalent complex between binding moieties means that melting temperature of the
35   oligonucleotide clamp incorporating the given pair(s) of binding moieties and its
target polynucleotide is increased by at least twenty-five percent over the melting
temperature of oligonucleotide moieties of the clamp alone, wherein melting
temperature is measured by standard techniques, e.g. half maximum of 260 nm

absorbance v. temperature as described more fully below. Preferably, stable means that melting temperature of the oligonucleotide clamp incorporating the given pair(s) of binding moieties and its target polynucleotide is increased by at least fifty percent over the melting temperature of oligonucleotide moieties of the clamp alone.

5

Detailed Description of the Invention

The invention provides a method of labeling and sorting molecules, particularly polynucleotides, by the use of oligonucleotide tags. In one aspect, the oligonucleotide tags of the invention comprise a plurality of "words" or subunits

10   selected from minimally cross-hybridizing sets of subunits. Subunits of such sets cannot form a duplex or triplex with the complement of another subunit of the same set with less than two mismatched nucleotides. Thus, the sequences of any two oligonucleotide tags of a repertoire that form duplexes will never be "closer" than differing by two nucleotides. In particular embodiments, sequences of any two

15   oligonucleotide tags of a repertoire can be even "further" apart, e.g. by designing a minimally cross-hybridizing set such that subunits cannot form a duplex with the complement of another subunit of the same set with less than three mismatched nucleotides, and so on. Usually, oligonucleotide tags of the invention and their complements are oligomers of the natural nucleotides so that they may be

20   conveniently processed by enzymes, such as ligases, polymerase, terminal transferases, and the like.

In another aspect of the invention, tag complements consist of monomers referred to herein as "antisense monomers." This term is meant to encompass a range of compounds typically developed for antisense therapeutics that have enhanced

25   binding strength and enhanced specificity for polynucleotide targets. As mentioned above under the definition of "oligonucleotide," the compounds may include a variety of different modifications of the natural nucleotides, e.g. modification of base moieties, sugar moieties, and/or monomer-to-monomer linkages. Such compounds also include oligonucleotide loops, oligonucleotide "clamps," and like structures,

30   described more fully below, that promote enhanced binding and specificity. Preferably, as used herein, the term "antisense monomer" means any monomer that can form a linear oligomer (i) capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, (ii) having a specificity of binding at least 25 percent higher than that of an equivalent

35   oligomer of natural nucleotides, and (iii) having a stability, or binding energy, equal to or greater than an equivalent oligomer of natural nucleotides, e.g. as measured by melting temperature of a duplex of an oligomer of antisense monomers and a complementary oligomer of natural nucleotides . "Specificity of binding" may be

measured in a variety of ways; preferably, it is measured by the difference in binding energy between a perfectly matched duplex and a duplex containing a single mismatch. Conveniently, melting temperature may be employed as a measure of binding energy. Clearly, the measure of specificity employed could be an average of

5   melting temperature differences of all possible mismatches of the test oligomer, it could be a minimum value or maximum value, or like function of melting temperature differences. Preferably, the minimal melting temperature difference between a perfectly matched duplex of an oligomer of antisense monomers and its complement, and a duplex of the same oligomer of antisense monomers and a

10   complement with a single mismatch is at least 25 percent greater than the equivalent melting temperature difference between duplexes of an oligomer of natural nucleotides and its perfectly matched and single-base-mismatched complements.

By the use of such compounds, the invention finds particular utility in labeling and sorting polynucleotides for parallel operations, such as sequencing, fingerprinting

15   or other types of analysis.

Constructing Oligonucleotide Tags from
Minimally Cross-Hybridizing Sets of Subunits

The nucleotide sequences of the subunits for any minimally cross-hybridizing

20   set are conveniently enumerated by simple computer programs following the general algorithm illustrated in Fig. 3, and as exemplified by program minhx whose source code is listed in Appendix I. Minhx computes all minimally cross-hybridizing sets having subunits composed of three kinds of nucleotides and having length of four.

The algorithm of Fig. 3 is implemented by first defining the characteristic of

25   the subunits of the minimally cross-hybridizing set, i.e. length, number of base differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table M_n, n=1, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit S_1 is selected and compared (120) with successive subunits S_i for i=n+1 to the end of the table.

30   Whenever a successive subunit has the required number of mismatches to be a member of the minimally cross-hybridizing set, it is saved in a new table M_{n+1} (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, M_2 will contain S_1; in the second set of comparisons, M_3 will contain S_1 and S_2; in the third set of comparisons, M_4 will

35   contain S_1, S_2, and S_3; and so on. Similarly, comparisons in table M_j will be between S_j and all successive subunits in M_j. Note that each successive table M_{n+1} is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table M_n has been compared (140) the old

table is replaced by the new table M_{n+1}, and the next round of comparisons are begun. The process stops (160) when a table M_n is reached that contains no successive subunits to compare to the selected subunit S_i, i.e. $M_n = M_{n+1}$.

      As mentioned above, preferred minimally cross-hybridizing sets comprise
5   subunits that make approximately equivalent contributions to duplex stability as every other subunit in the set. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit.
10   Rev. Biochem. Mol. Biol., 26: 227-259 (1991);and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed.
15   A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5' →3' exonuclease activity of a DNA polymerase. The following is an exemplary
20   minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

<div align="center">

Table I

Word:	w_1	w_2	w_3	w_4
Sequence:	GATT	TGAT	TAGA	TTTG

Word:	w_5	w_6	w_7	w_8
Sequence:	GTAA	AGTA	ATGT	AAAG

</div>

25

In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

      Further exemplary minimally cross-hybridizing sets are listed below in Table
30   I. Clearly, additional sets can be generated by substituting different groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

Table II

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

5   The oligonucleotide tags of the invention and their complements are
conveniently synthesized on an automated DNA synthesizer, e.g. an Applied
Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer,
using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the
following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko
10   et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S.
patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries,
e.g. resulting in non-natural backbone groups, such as phosphorothioate,
phosphoramidate, and the like, may also be employed provided that the resulting
oligonucleotides are capable of specific hybridization. In some embodiments, tags
15   may comprise naturally occuring nucleotides that permit processing or manipulation
by enzymes, while the corresponding tag complements may comprise non-natural
nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the
formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags
20   and tag complements are preferably generated by subunit-wise synthesis via "split and
mix" techniques, e.g. as disclosed in Shortle et al, International patent application
PCT/US93/03418. Briefly, the basic unit of the synthesis is a subunit of the

oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct

5   analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J.

10   Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of mixtures of the activated monomers to the growing oligonucleotide during the coupling steps.

Double stranded forms of tags are made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex

15   formation. Such duplex tags may then be inserted into cloning vectors along with target polynucleotides for sorting and manipulation of the target polynucleotide in accordance with the invention.

In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags;

20   however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*" indicates Hoogsteen type of binding); however, other motifs are also possible. For

25   example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (e.g. whether ribose or

30   deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991);

35   Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan, Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem.,

267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992);
Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene,
Angew. Chem. Int. Ed. Engl. 32: 666-690 (1993); and the like. Conditions for
annealing single-stranded or duplex tags to their single-stranded or duplex
5   complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-1328 (1993).

When oligomeric subunits are employed, oligonucleotide tags of the invention
and their complements may range in length from 12 to 60 nucleotides or basepairs;
more preferably, they range in length from 18 to 40 nucleotides or basepairs; and
most preferably, they range in length from 25 to 40 nucleotides or basepairs. When
10   constructed from antisense monomers, oligonucleotide tags and their complements
preferably range in length from 10 to 40 monomers; and more preferably, they range
in length from 12 to 30 monomers.

<div align="center">

Table III
Numbers of Subunits in Tags in Preferred Embodiments

</div>

15

Monomers in Subunit	Nucleotides in Oligonucleotide Tag		
	(12-60)	(18-40)	(25-40)
3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

20   Most preferably, oligonucleotide tags are single stranded and specific hybridization
occurs via Watson-Crick pairing with a tag complement.

<div align="center">

Solid Phase Supports

</div>

Solid phase supports for use with the invention may have a wide variety of
25   forms, including microparticles, beads, and membranes, slides, plates, micromachined
chips, and the like. Likewise, solid phase supports of the invention may comprise a
wide variety of compositions, including glass, plastic, silicon, alkanethiolate-
derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica
gel, polyamide, and the like. Preferably, either a population of discrete particles are
30   employed such that each has a uniform coating, or population, of complementary
sequences of the same tag (and no other), or a single or a few supports are employed

with spacially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several μm^2, e.g. 3-5, to several hundred μm^2, e.g. 100-

5   500. Preferably, such regions are spacially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

10   Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, Nucleic Acids Research, 16: 10861-10880 (1988); Albretsen et al, Anal. Biochem., 189: 40-50 (1990); Wolf et al, Nucleic Acids Research, 15: 2911-2926 (1987); or Ghosh et al, Nucleic Acids Research, 15:

15   5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including

20   microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: Meth. Enzymol., Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046;720; and Pon, Chapter 19, in Agrawal, editor, Methods in

25   Molecular Biology, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel™, Rapp Polymere, Tubingen Germany); and the like. Selection of the

30   support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hinderance of the enzymes and that facilitate access to substrate are preferred. Exemplary linking moieties are disclosed in Pon et

35   al, Biotechniques, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, J. Chem. Soc. Commun., 1989: 891-893; Damha et al, Nucleic Acids Research, 18: 3813-3821

(1990); Beattie et al, Clinical Chemistry, 39: 719-722 (1993); Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag
5   complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, Genomics, 13:
10   1008-1017 (1992); and Maskos and Southern, Nucleic Acids Research, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces
15   are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, Oligonucleotide Synthesis: A Practical Approach (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the
20   construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

Preferably, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached, e.g. Applied
25   Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large numbers of microparticles on planar supports, such as a microscope slide.
30   Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μm diameter GMA beads are employed.
35

Synthesis of Oligonucleotide 3'N→5'P Phosphoramidates

Tag complements comprising oligonucleotide 3'N→5'P phosphoramidates may be synthesized on a solid support using the step-by-step elongation procedure

- 14 -

outlined in Figure 4. The synthetic cycle for addition of a single aminonucleoside consists essentially of the following operations: detritylation (step a); phosphitylation of the 5' hydroxyl group to generate a 5'-hydrogen phosphonate diester (steps b and c); and Atherton-Todd type coupling of a 5'-DMT-3'-aminonucleoside (e.g. as

5   disclosed by Glinski et al, Chem. Comm., pp. 915-916 (1970)) with the 5' hydrogen phosphonate in the presence of carbon tetrachloride (step d). Coupling yields range between 94-96% per cycle. The resulting oligonucleotide phosphoramidate is cleaved and deprotected with ammonia and thereafter purified by ion exchange high performance liquid chromatography. The following references provide guidance for

10   carrying out the above synthesis: Atherton et al, J. Chem. Soc., pp. 660-663 (1945); Gryaznov et al, Nucleic Acids Research, 20: 3403-3409 (1992); Gryaznov et al, Vest. Mosk. Univ. Ser. 2: Khim 27: 421-424 (1986); and Gryaznov et al, Tetrahedron Lett., 31: 3205-3208 (1990). Fung and Gryaznov, International application PCT/US94/03087, also show that 3'-amino-oligonucleotides may be enzymatically

15   ligated to 5'-phosphorylated oligonucleotides in a standard template-driven ligation reaction.

Oligonucleotide Clamps

Tag complements of the invention may comprise an oligonucleotide clamp, which is a compound capable of forming a covalently closed macrocycle or a stable

20   circular complex after specifically binding to a target polynucleotide, which in the case of the present invention is its corresponding oligonucleotide tag. Generally, oligonucleotide clamps comprise one or more oligonucleotide moieties capable of specifically binding to a tag and one or more pairs of binding moieties covalently linked to the oligonucleotide moieties. Upon annealing of the oligonucleotide

25   moieties to the target polynucleotide, the binding moieties of a pair are brought into juxtaposition so that they form a stable covalent or non-covalent linkage or complex. The interaction of the binding moieties of the one or more pairs effectively clamps the specifically annealed oligonucleotide moieties to the target polynucleotide.

In one preferred form oligonucleotide clamps comprise a first binding moiety,

30   a first oligonucleotide moiety, a hinge region, a second oligonucleotide moiety, and a second binding moiety, for example, as represented by the particular embodiment of the following formula:

$$X\text{-}O_1\text{-}G\text{-}O_2\text{-}Y$$

35

wherein O_1 and O_2 are the first and second oligonucleotide moieties, G is the hinge region, X is the first binding moiety and Y is the second binding moiety such that X and Y form a stable covalent or non-covalent linkage or complex whenever they are

brought into juxtapositon by the annealing of the oligonucleotide moieties to a target polynucleotide. Preferably, in this embodiment, one of O_1 and O_2 undergoes Watson-Crick binding with the target polynucleotide while the other of O_1 and O_2 undergoes Hoogsteen binding.

5   Preferably, stability of oligonucleotide clamp/target polynucleotide complexes are determined by way of melting, or strand dissociation, curves. The temperature of fifty percent strand dissociation is taken as the melting temperature, T_m, which, in turn, provides a convenient measure of stability. T_m measurements are typically carried out in a saline solution at neutral pH with target and clamp concentrations at

10   between about 1.0-2.0 μM. Typical conditions are as follows: 150 mM NaCl and 10 mM $MgCl_2$ in a 10 mM sodium phosphate buffer (pH 7.0) or in a 10 mM Tris-HCl buffer (pH 7.0); or like conditions. Data for melting curves are accumulated by heating a sample of the oligonucleotide clamp/target polynucleotide complex from room temperature to about 85-90°C. As the temperature of the sample increases,

15   absorbance of 260 nm light is monitored at 1°C intervals, e.g. using a Cary (Australia) model 1E or a Hewlett-Packard (Palo Alto, CA) model HP 8459 UV/VIS spectrophotometer and model HP 89100A temperature controller, or like instruments.
          Hinge regions consist of nucleosidic or non-nucleosidic polymers which preferably facilitate the specific binding of the monomers of the oligonucleotide

20   moieties with their complementary nucleotides of the target polynucleotide. Hinge regions may also include linkages to solid phase supports, e.g. via a derivatized base of a nucleotide, or the like. Generally, the oligonucleotide moieties may be connected to hinge regions and/or binding moieties in either 5'→3' or 3'→5' orientations. For example, in the embodiment described above comprising a first binding moiety, a

25   first oligonucleotide moiety, a hinge region, a second oligonucleotide moiety, and a second binding moiety, the oligonucleotide moieties may have any of the following orientations:

$$X-(5')N_1N_2N_3- ... -N_j(3')-G-(5')N_1N_2N_3- ... -N_k(3')-Y$$

30   OR

$$X-(5')N_1N_2N_3- ... -N_j(3')-G-(3')N_kN_{k-1}N_{k-2}- ... -N_1(5')-Y$$

    OR

$$X-(3')N_jN_{j-1}N_{j-2}- ... -N_1(5')-G-(5')N_1N_2N_3- ... -N_k(3')-Y$$

    OR

35   $$X-(3')N_jN_{j-1}N_{j-2}- ... -N_1(5')-G-(3')N_kN_{k-1}N_{k-2}- ... -N_1(5')-Y$$

wherein $N_1N_2N_3- ... -N_k$ and $N_1N_2N_3- ... -N_j$ are k-mer and j-mer oligonucleotide moieties in the indicated orientations.

Preferably, the hinge region is a linear oligomer of monomers selected from the group consisting of alkyl, alkenyl, and/or ethers containing 2-3 carbon atoms. Preferably, for nucleoside-sized monomers or smaller, the number of monomers varies between about 3 and about 10; and more preferably, it varies between about 4 and 8.

A variety of binding moieties are suitable for use with the invention. Generally, they are employed in pairs, which for convenience here will be referred to as X and Y. X and Y may be the same or different. Whenever the interaction of X and Y is based on the formation of stable hydrophobic complex, X and Y are lipophilic groups, including alkyl groups, fatty acids, fatty alcohols, steroids, waxes, fat-soluble vitamins, and the like. Further exemplary lipophilic binding moieties include glycerides, glyceryl ethers, phospholipids, sphingolipids, terpenes, and the like. In such embodiments, X and Y are preferably selected from the group of steroids consisting of a derivatized perhydrocyclopentanophenanthrene nucleus having from 19 to 30 carbon atoms, and 0 to 6 oxygen atoms; alkyl having from 6 to 16 carbon atoms; vitamin E; and glyceride having 20 to 40 carbon atoms. Preferably, a perhydrocyclopentanophenanthrene-based moiety is attached through the hydroxyl group, either as an ether or an ester, at its C3 position. It is understood tha X and Y may include a linkage group connecting it to an oligonucleotide moiety. For example, glyceride includes phosphoglyceride, e.g. as described by MacKellar et al, Nucleic Acids Research, 20: 3411-3417 (1992), and so on. It is especially preferred that lipophilic moieties, such as perhydrocyclopentanophenanthrene derivatives, be linked to the 5' carbon and/or the 3' carbon of an oligonucleotide moiety by a short but flexible linker that permits the lipophilic moiety to interact with the bases of the oligonucleotide clamp/target polynucleotide complex or a lipophilic moiety on the same or another oligonucleotide moiety. Such linkers include phosphate (i.e. phosphodiester), phosphoramidate, hydroxyurethane, carboxyaminoalkyl and carboxyaminoalkylphosphate linkers, or the like. Preferably, such linkers have no more than from 2 to 8 carbon atoms.

Binding moieties can be attached to the oligonucleotide moiety by a number of available chemistries. Generally, it is preferred that the oligonucleotide be initially derivatized at its 3' and/or 5' terminus with a reactive functionality, such as an amino, phosphate, thiophosphate, or thiol group. After derivatization, a hydrophilic or hydrophobic moiety is coupled to the oligonucleotide via the reactive functionality. Exemplary means for attaching 3' or 5' reactive functionalities to oligonucleotides are disclosed in Fung et al, U.S. patent 5,212,304; Connolly, Nucleic Acids Research, 13: 4485-4502 (1985); Tino, International application PCT/US91/09657; Nelson et al, Nucleic Acids Research, 17: 7187-7194 (1989); Stabinsky, U.S. patent 4,739,044;

Gupta et al, Nucleic Acids Research, 19: 3019 (1991); Reed et al, International application PCT/US91/06143; Zuckerman et al, Nucleic Acids Research, 15: 5305 (1987); Eckstein, editor, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991); Clontech 1992/1993 Catalog (Clontech Laboratories, Palo Alto,
5   CA); and like references.

Preferably, whenever X and Y form a covalent linkage, X and Y pairs must react specifically with each other when brought into juxtaposition. In this aspect of the invention, X and Y pairs are preferably selected from the following group: when one of X or Y is phosphorothioate or phosphorodithioate, the other is haloacetyl,
10   haloacyl, haloalkyl, or alkylazide; when one of X or Y is thiol, the other is alkyl iodide, haloacyl, or haloacetyl; when one of Y or Y is phenylazide the other is phenylazide. More preferably, when one of X or Y is phosphorothioate or phosphorodithioate, the other is haloacetyl, haloacyl, or haloalkyl, wherein said alkyl, acetyl, or acyl moiety contains from one to eight carbon atoms. Such chemistries are
15   disclosed by Gryaznov et al, J. Am. Chem. Soc., 115: 3808-3809 (1993).

In some embodiments, X and Y may form a covalent linkage in the presence of an activating agent. That is, one or both of the binding moieties are activated or rendered reactive towards one another by exposure to an activating agent or condensing agent, such as radiation, a reducing agent, an oxidizing agent, or the like.
20   Exemplary, binding moieties employing activating agents include thiophosphoryl groups in the presence of $K_3Fe(CN)_6$ or KI_3, e.g. Gryaznov and Letsinger, Nucleic Acids Research, 21: 1403-1408 (1993); phosphoryl and hydroxyl in the presence of N-cyanoimidazole, e.g. Luebke et al, J. Am. Chem. Soc., 113: 7447-7448 (1991); phosphoryl or amino group and hydroxyl in the presence of cyanogen bromide, e.g.
25   Sokolova et al, FEBS Letters, 232: 153-155 (1988); phosphoryl and hydroxyl groups in the presence of spermine-5-(N-ethylimidazole)carboxamide and cyanoimidazole, e.g. Zuber et al, J. Am. Chem. Soc., 115: 4939-4940 (1993); and the like.

Attaching Target Polynucleotides to Microparticles

30   An important aspect of the invention is the sorting of populations of identical polynucleotides, e.g. from a cDNA library, and their attachment to microparticles or separate regions of a solid phase support such that each microparticle or region has only a single kind of polynucleotide. This latter condition can be essentially met by ligating a repertoire of tags to a population of polynucleotides followed by cloning
35   and sampling of the ligated sequences. A repertoire of oligonucleotide tags can be ligated to a population of polynucleotides in a number of ways, such as through direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of

tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. (Note that it is also possible to obtain

5   different tags with the same polynucleotide in a sample. This case is simply leads to a polynucleotide being processed, e.g. sequenced, twice). As explain more fully below, the probability of obtaining a double in a sample can be estimated by a Poisson distribution since the number of conjugates in a sample will be large, e.g. on the order of thousands or more, and the probability of selecting a particular tag will be small

10   because the tag repertoire is large, e.g. on the order of tens of thousand or more. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation, and selecting a small sample which ensures that a

15   minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored. As used herein, the term "substantially all" in reference to attaching tags to molecules, especially

20   polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the tags

25   have unique polynucleotides attached. More preferably, it means that at least ninety percent of the tags have unique polynucleotides attached. Still more preferably, it means that at least ninety-five percent of the tags have unique polynucleotides attached. And, most preferably, it means that at least ninety-nine percent of the tags have unique polynucleotides attached.

30   Preferably, when the population of polynucleotides is messenger RNA (mRNA), oligonucleotides tags are attached by reverse transcribing the mRNA with a set of primers containing complements of tag sequences. An exemplary set of such primers could have the following sequence:

35            $5'$-mRNA- $[A]_n$ -$3'$
                     $[T]_{19}GG[W,W,W,C]_9AC\underline{CAGCTGA}TC$-$5'$-biotin

where "[W,W,W,C]9" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

$$5'-[G,W,W,W]_9\text{TGG-linker-microparticle}$$

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion, and the second strand of the cDNA is synthesized using, for example, a primer of the following form:

$$5'-\text{NRRGATCYNNN-3}'$$

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

$$5'-\text{RCGACCA}[C,W,W,W]_9\text{GG}[T]_{19}-\text{cDNA}\quad-\text{NNNR}$$
$$\text{GGT}[G,W,W,W]_9\text{CC}[A]_{19}-\text{rDNA}\quad-\text{NNNYCTAG-5}'$$

Preferably, when the ligase-based method of sequencing is employed, the Bst Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites:

$$5'-\text{GA}\underline{\text{GGATG}}\text{CCTTTAT}\underline{\text{GGATCC}}\text{A}\underline{\text{CTCGAG}}\text{ATCCCAATCCA-3}'$$
$$\quad\quad\quad\text{FokI}\quad\quad\quad\quad\text{BamHI}\quad\text{XhoI}$$

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'→3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single nucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double

stranded fragment until a complement of the single nucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the 5'→3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, tags constructed

5   with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double standed portion of

10   the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed

15   with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et

20   al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags are ligated to the complementary sequences attached to the microparticles.

25   Finally, the microparticles are washed to remove unligated polynucleotides.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a

30   variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so

35   high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, Nucleic Acids Research, 20: 1679-1684

(1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range of 20-50 μm are loaded with about 10^5 polynucleotides.

The above method may be used to fingerprint mRNA populations when coupled with the parallel sequencing methodology described below. Partial sequence

5   information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, of cDNAs attached to separate microparticles as described in the above method. The frequency distribution of partial sequences can identify mRNA populations from different cell or tissue types, as well as from diseased tissues, such as cancers. Such mRNA fingerprints are useful in monitoring and diagnosing disease

10   states, e.g. International application PCT/US95/21944, which describes the use of express sequence tags (ESTs) for the same purpose.

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA

15   sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably,

20   the methodology permits the stepwise identification of nucleotides, usually one at a time, in a sequence through successive cycles of treatment and detection. Such methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al,

25   International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

A "single base" method of DNA sequencing which is suitable for use with the present invention and whch requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. The method comprises

30   the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by

35   the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide is determined. As is described more fully below, identifying the one or more nucleotides can be carried out either before or after cleavage of the ligated complex

from the target polynucleotide. Preferably, whenever natural protein endonucleases are employed, the method further includes a step of methylating the target polynucleotide at the start of a sequencing operation.

An important feature of the method is the probe ligated to the target
5   polynucleotide. A preferred form of the probes is illustrated in Figure 2. Generally, the probes are double stranded DNA with a protruding strand at one end 10. The probes contain at least one nuclease recognition site 12 and a spacer region 14 between the recognition site and the protruding end 10. Preferably, probes also include a label 16, which in this particular embodiment is illustrated at the end
10   opposite of the protruding strand. The probes may be labeled by a variety of means and at a variety of locations, the only restriction being that the labeling means selected does not interfer with the ligation step or with the recognition of the probe by the nuclease.

It is not critical whether protruding strand 10 of the probe is a 5' or 3' end.
15   However, it is important that the protruding strands of the target polynucleotide and probes be capable of forming perfectly matched duplexes to allow for specific ligation. If the protruding strands of the target polynucleotide and probe are different lengths the resulting gap can be filled in by a polymerase prior to ligation, e.g. as in "gap LCR" disclosed in Backman et al, European patent application 91100959.5.
20   Preferably, the number of nucleotides in the respective protruding strands are the same so that both strands of the probe and target polynucleotide are capable of being ligated without a filling step. Preferably, the protruding strand of the probe is from 2 to 6 nucleotides long. As indicated below, the greater the length of the protruding strand, the greater the complexity of the probe mixture that is applied to the target
25   polynucleotide during each ligation and cleavage cycle.

The complementary strands of the probes are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries. After synthesis, the complementary strands are combined to form a double stranded
30   probe. Generally, the protruding strand of a probe is synthesized as a mixture, so that every possible sequence is represented in the protruding portion. For example, if the protruding portion consisted of four nucleotides, in one embodiment four mixtures are prepared as follows:

$$X_1X_2 \ldots X_i\text{NNNA},$$
35   $$X_1X_2 \ldots X_i\text{NNNC},$$
$$X_1X_2 \ldots X_i\text{NNNG, and}$$
$$X_1X_2 \ldots X_i\text{NNNT}$$

where the "NNNs" represent every possible 3-mer and the "Xs" represent the duplex forming portion of the strand. Thus, each of the four probes listed above contains 4^3 or 64 distinct sequences; or, in other words, each of the four probes has a degeneracy of 64. For example, $X_1X_2 \ldots X_iNNNA$ contains the following sequences:

5

$$X_1X_2 \ldots X_iAAAA$$
$$X_1X_2 \ldots X_iAACA$$
$$X_1X_2 \ldots X_iAAGA$$
$$X_1X_2 \ldots X_iAATA$$
10
$$X_1X_2 \ldots X_iACAA$$

.

.

.

15

$$X_1X_2 \ldots X_iTGTA$$
$$X_1X_2 \ldots X_iTTAA$$
$$X_1X_2 \ldots X_iTTCA$$
$$X_1X_2 \ldots X_iTTGA$$
$$X_1X_2 \ldots X_iTTTA$$
20

Such mixtures are readily synthesized using well known techniques, e.g. as disclosed in Telenius et al (cited above). Generally, these techniques simply call for the application of mixtures of the activated monomers to the growing oligonucleotide during the coupling steps where one desires to introduce the degeneracy. In some

25   embodiments it may be desirable to reduce the degeneracy of the probes. This can be accomplished using degeneracy reducing analogs, such as deoxyinosine, 2-aminopurine, or the like, e.g. as taught in Kong Thoo Lin et al, Nucleic Acids Research, 20: 5149-5152, or by U.S. patent 5,002,867.

Preferably, for oligonucleotides with phosphodiester linkages, the duplex

30   forming region of a probe is between about 12 to about 30 basepairs in length; more preferably, its length is between about 15 to about 25 basepairs.

When conventional ligases are employed in the invention, as described more fully below, the 5' end of the probe may be phosphorylated in some embodiments. A 5' monophosphate can be attached to a second oligonucleotide either chemically or

35   enzymatically with a kinase, e.g. Sambrook et al (cited above). Chemical phosphorylation is described by Horn and Urdea, Tetrahedron Lett., 27: 4705 (1986), and reagents for carrying out the disclosed protocols are commercially available, e.g.

5' Phosphate-ON^{TM} from Clontech Laboratories (Palo Alto, California). Thus, in some embodiments, probes may have the form:

$$5'\text{-}X_1X_2 \ldots X_i\text{TTGA}$$
$$Y_1Y_2 \ldots Y_ip$$

where the Y's are the complementary nucleotides of the X's and "p" is a monophosphate group.

The above probes can be labeled in a variety of ways, including the direct or indirect attachment of radioactive moieties, fluorescent moieties, colorimetric moieties, chemiluminescent markers, and the like. Many comprehensive reviews of methodologies for labeling DNA and constructing DNA probes provide guidance applicable to constructing probes of the present invention. Such reviews include Kricka, editor, Nonisotopic DNA Probe Techniques (Academic Press, San Diego, 1992); Haugland, Handbook of Fluorescent Probes and Research Chemicals (Molecular Probes, Inc., Eugene, 1992); Keller and Manak, DNA Probes, 2nd Edition (Stockton Press, New York, 1993); and Eckstein, editor, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991); Kessler, editor, Nonradioactive Labeling and Detection of Biomolecules (Springer-Verlag, Berlin, 1992); Wetmur (cited above); and the like.

Preferably, the probes are labeled with one or more fluorescent dyes, e.g. as disclosed by Menchen et al, U.S. patent 5,188,934; Begot et al International application PCT/US90/05565.

In accordance with the method, a probe is ligated to an end of a target polynucleotide to form a ligated complex in each cycle of ligation and cleavage. The ligated complex is the double stranded structure formed after the protruding strands of the target polynucleotide and probe anneal and at least one pair of the identically oriented strands of the probe and target are ligated, i.e. are caused to be covalently linked to one another. Ligation can be accomplished either enzymatically or chemically. Chemical ligation methods are well known in the art, e.g. Ferris et al, Nucleosides & Nucleotides, 8: 407-414 (1989); Shabarova et al, Nucleic Acids Research, 19: 4247-4251 (1991); and the like. Preferably, however, ligation is carried out enzymatically using a ligase in a standard protocol. Many ligases are known and are suitable for use in the invention, e.g. Lehman, Science, 186: 790-797 (1974); Engler et al, DNA Ligases, pages 3-30 in Boyer, editor, The Enzymes, Vol. 15B (Academic Press, New York, 1982); and the like. Preferred ligases include T4 DNA ligase, T7 DNA ligase, E. coli DNA ligase, Taq ligase, Pfu ligase, and Tth ligase. Protocols for their use are well known, e.g. Sambrook et al (cited above); Barany, PCR Methods and Applications, 1: 5-16 (1991); Marsh et al, Strategies, 5: 73-76

(1992); and the like. Generally, ligases require that a 5' phosphate group be present for ligation to the 3' hydroxyl of an abutting strand. This is conveniently provided for at least one strand of the target polynucleotide by selecting a nuclease which leaves a 5' phosphate, e.g. as Fok I.

5   In an embodiment of the sequencing method employing unphosphorylated probes, the step of ligating includes (i) ligating the probe to the target polynucleotide with a ligase so that a ligated complex is formed having a nick on one strand, (ii) phosphorylating the 5' hydroxyl at the nick with a kinase using conventional protocols, e.g. Sambrook et al (cited above), and (iii) ligating again to covalently join

10   the strands at the nick, i.e. to remove the nick.

Apparatus for Observing Enzymatic Processes and/or Binding Events at Microparticle Surfaces

An objective of the invention is to sort identical molecules, particularly

15   polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like.

20   Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded"

25   microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with

30   a scanning system, such as described in International patent applications PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional

35   identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a

CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with

5   fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available form Apple Computer (Cupertino, CA). Computer software for table translation and data collection functions can be provided

10   by commercially available laboratory software, such as Lab Windows, available from National Instruments.

     Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from

15   light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or

20   more suitable opto-electronic devices for converting some fluorescence characteristic, e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 5, signal distribution optics 318 directs

25   fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336, whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD

30   array.

     The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for

35   most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100 μm. Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a

compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of

5 fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per cm^2.

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without

10 significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning

15 vectors used to attach tags to polynucleotides are engineered to contain a unique restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique

20 restriction site distal from the microparticle surface. After digestion with the associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

25 Alternatively and preferably when sequencing by ligation is employed, in the initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type IIs restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture

30 comprises about 10-15 percent of the biotinylated probe.

In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles.

35 Preferably, such a glass substrate is a flow cell, which may comprise a channel etched in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

Parallel Sequencing

The tagging system of the invention can be used with single base sequencing methods to sequence polynucleotides up to several kilobases in length. The tagging

5   system permits many thousands of fragments of a target polynucleotide to be sorted onto one or more solid phase supports and sequenced simultaneously. In accordance with a preferred implementation of the method, a portion of each sorted fragment is sequenced in a stepwise fashion on each of the many thousands of loaded microparticles which are fixed to a common substrate—such as a microscope slide—

10   associated with a scanning system or an image analysis system, such as described above. The size of the portion of the fragments sequenced depends of several factors, such as the number of fragments generated and sorted, the length of the target polynucleotide, the speed and accuracy of the single base method employed, the number of microparticles and/or discrete regions that may be monitored

15   simultaneously; and the like. Preferably, from 12-50 bases are identified at each microparticle or region; and more preferably, 18-30 bases are identified at each microparticle or region. With this information, the sequence of the target polynucleotide is determined by collating the 12-50 base fragments via their overlapping regions, e.g. as described in U.S. patent 5,002,867. The following

20   references provide additional guidance in determining the portion of the fragments that must be sequenced for successful reconstruction of a target polynucleotide of a given length: Lander and Waterman, Genomics, 2: 231-239 (1988); Drmanac et al, Genomics, 4: 114-128 (1989); Bains, DNA Sequencing and Mapping, 4: 143-150 (1993); Bains, Genomics, 11: 294-301 (1991); Drmanac et al, J. Biomolecular

25   Structure and Dynamics, 8: 1085-1102 (1991); and Pevzner, J. Biomolecular Structure and Dynamics, 7: 63-73 (1989). Preferably, the length of the target polynucleotide is between 1 kilobase and 50 kilobases. More preferably, the length is between 10 kilobases and 40 kilobases. Lander and Waterman (cited above) provide guidance concerning the relationship among the number of fragments that are

30   sequenced (i.e. the sample size), the amount of sequence information obtained from each fragment, and the probability that the target polynucleotide can be reconstructed from the partial sequences without gaps, or "islands." For the present invention, maximal polynucleotide sizes that can be obtained for given sample sizes and sizes of fragment sequences are shown below:

35

Size of Sample	Approx. maximal target polynucleotide length	
	30 bases/fragment	50 bases/fragment
1,000	3 kilobases	4 kilobases
10,000	22 kilobases	32 kilobases
20,000	40 kilobases	65 kilobases
30,000	60 kilobases	85 kilobases
100,000	180 kilobases	300 kilobases

Fragments may be generated from a target polynucleotide in a variety of ways, including so-called "directed" approaches where one attempts to generate sets of
5   fragments covering the target polynucleotide with minimal overlap, and so-called "shotgun" approaches where randomly overlapping fragments are generated. Preferably, "shotgun" approaches to fragment generation are employed because of their simplicity and inherent redundancy. For example, randomly overlapping fragments that cover a target polynucleotide are generated in the following
10   conventional "shotgun" sequencing protocol, e.g. as disclosed in Sambrook et al (cited above). As used herein, "cover" in this context means that every portion of the target polynucleotide sequence is represented in each size range, e.g. all fragments between 100 and 200 basepairs in length, of the generated fragments. Briefly, starting with a target polynucleotide as an insert in an appropriate cloning vector, e.g. λ
15   phage, the vector is expanded, purified and digested with the appropriate restriction enzymes to yield about 10-15 μg of purified insert. Typically, the protocol results in about 500-1000 subclones per microgram of starting DNA. The insert is separated from the vector fragments by preparative gel electrophoresis, removed from the gel by conventional methods, and resuspended in a standard buffer, such as TE (Tris-
20   EDTA). The restriction enzymes selected to excise the insert from the vector preferably leave compatible sticky ends on the insert, so that the insert can be self-ligated in preparation for generating randomly overlapping fragments. As explained in Sambrook et al (cited above), the circularized DNA yields a better random distribution of fragments than linear DNA in the fragmentation methods employed
25   below. After self-ligating the insert, e.g. with T4 ligase using conventional protocols, the purified ligated insert is fragmented by a standard protocol, e.g. sonication or DNase I digestion in the presence of Mn^{++}. After fragmentation the ends of the fragments are repair, e.g. as described in Sambrook et al (cited above), and the repaired fragments are separated by size using gel electrophoresis. Fragments in the
30   300-500 basepair range are selected and eluted from the gel by conventional means,

and ligated into a tag-carrying vector as described above to form a library of tag-fragment conjugates.

As described above, a sample containing several thousand tag-fragment conjugates are taken from the library and expanded, after which the tag-fragment inserts are excised from the vector and prepared for specific hybridization to the tag complements on microparticles, as described above. Depending of the size of the target polynucleotide, multiple samples may be taken from the tag-fragment library and separately expanded, loaded onto microparticles and sequenced. The number of doubles selected will depend on the fraction of the tag repertoire represented in a sample. (The probability of obtaining triples--three different polynucleotides with the same tag-- or above can safely be ignored). As mentioned above, the probability of doubles in a sample can be estimated from the Poisson distribution $p(double)=m^2e^{-m}/2$, where m is the fraction of the tag repertoire in the sample. Table V below lists probabilities of obtaining doubles in a sample for given tag size, sample size, and repertoire diversity.

Table V

Number of words in tag from 8 word set	Size of tag repertoire	Size of sample	Fraction of repertoire sampled	Probability of double
7	2.1×10^6	3000	1.43×10^{-3}	10^{-6}
8	1.68×10^7	3×10^4	1.78×10^{-3}	1.6×10^{-6}
		3000	1.78×10^{-4}	1.6×10^{-8}
9	1.34×10^8	3×10^5	2.24×10^{-3}	2.5×10^{-6}
		3×10^4	2.24×10^{-4}	2.5×10^{-8}
10	1.07×10^9	3×10^6	2.8×10^{-3}	3.9×10^{-6}
		3×10^5	2.8×10^{-4}	3.9×10^{-8}

In any case, the loaded microparticles are then dispersed and fixed onto a glass microscope slide, preferably via an avidin-biotin coupling. Preferably, at least 15-20 nucleotides of each of the random fragments are simultaneously sequenced with a single base method. The sequence of the target polynucleotide is then reconstructed by collating the partial sequences of the random fragments by way of their overlapping portions, using algorithms similar to those used for assembling contigs, or as developed for sequencing by hybridization, disclosed in the above references.

Kits for Implementing the Method of the Invention

The invention includes kits for carrying out the various embodiments of the invention. Preferably, kits of the invention include a repertoire of tag complements

attached to a solid phase support. Additionally, kits of the invention may include the corresponding repertoire of tags, e.g. as primers for amplifying the polynucleotides to be sorted or as elements of cloning vectors which can also be used to amplify the polynucleotides to be sorted. Preferably, the repertoire of tag complements are

5   attached to microparticles. Kits may also contain appropriate buffers for enzymatic processing, detection chemistries, e.g. fluorescent or chemiluminescent tags, and the like, instructions for use, processing enzymes, such as ligases, polymerases, transferases, and so on. In an important embodiment for sequencing, kits may also include substrates, such as a avidinated microscope slides, for fixing loaded

10   microparticles for processing.

Identification of Novel Polynucleotides
in cDNA Libraries

Novel polynucleotides in a cDNA library can be identified by constructing a

15   library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may by hybridized to the polyA tails for first strand

20   synthesis with a reverse transcriptase using conventional protocols:

$$5'\text{-mRNA-}[A]_n\ \text{-}3'$$
$$[T]_{19}\text{-[primer site]-GG}[W,W,W,C]_9\text{ACCAGCTGATC-}5'$$

25   where $[W,W,W,C]_9$ represents a tag as described above, "ACCAGCTGATC" is an optional sequence forming a restriction site in double stranded form, and "primer site" is a sequence common to all members of the library that is later used as a primer binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional

30   techniques, the double stranded fragments are inserted into a cloning vector as described above and amplified. The amplified library is then sampled and the sample amplified. The cloning vectors from the amplified sample are isolated, and the tagged cDNA fragments excised and purified. After rendering the tag single stranded with a polymerase as described above, the fragments are methylated and sorted onto

35   microparticles in accordance with the invention. Preferably, as described above, the cloning vector is constructed so that the tagged cDNAs can be excised with an endonuclease, such as Fok I, that will allow immediate sequencing by the preferred single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library, or one or more large fractions of the library, in accordance with the invention until a sufficient number of nucleotides are identified on each cDNA for unique representation in the genome of the organism from which the library is derived. For

5   example, if the library is derived from mammalian mRNA then a randomly selected sequence 14-15 nucleotides long is expected to have unique representation among the 2-3 thousand megabases of the typical mammalian genome. Of course identification of far fewer nucleotides would be sufficient for unique representation in a library derived from bacteria, or other lower organisms. Preferably, at least 20-30

10   nucleotides are identified to ensure unique representation and to permit construction of a suitable primer as described below. The tabulated sequences may then be compared to known sequences to identify unique cDNAs.

Unique cDNAs are then isolated by conventional techniques, e.g. constructing a probe from the PCR amplicon produced with primers directed to the primer site and

15   the portion of the cDNA whose sequence was determined. The probe may then be used to identify the cDNA in a library using a conventional screening protocol.

Cycle Sequencing on Microparticles Loaded
with Sorted Polynucleotides

20   Parallel sequencing may also be accomplished in accordance with the invention with conventional sequencing techniques that require the generation and separation of labeled DNA fragments. In particular, isolated microparticles loaded with a uniform population of templates may be used to generate labeled extension products by cycle sequencing. Cycle sequencing is a well-know variant of the basic

25   Sanger approach to DNA sequencing describe fully in the following references: Craxton, Methods, Vol. 2 (February, 1991); Wozny, European patent publication 0 409 078 A2 (23 January 1991); Fuller, International application PCT/US92/07303; and Fuller, International application PCT/US94/03264. Briefly, in a standard sequencing reaction mixture, a thermal stable polymerase is employed so that

30   repeated extension reactions may be carried out on the same template. This permits small amounts of template to generate sufficient amounts of extension product for detection after separation by electrophoresis. Typically, cycle sequencing comprises the steps of (a) providing a sequencing reaction mixture with a template, a primer, nucleoside triphosphates, chain-terminating nucleoside triphosphates, and a thermal

35   stable DNA polymerase; (b) denaturing the template, (c) annealing the primer to the denatured template, (d) extending the primer to form extension products, and (e) repeating steps (b)-(d) until sufficient quantities of extension products are accumulated so that they may be detected upon separation. The number of times the

cycle is repeated depends on many factors, including the amount and quality of
starting template, the detection system employed, the separation system employed,
and the like. As conventionally practiced, the extension cycle is typically repeated
from 10 to 100 times; the template amount ranges from as little as a few tens of

5   femtomole to several tens of picomole; the denaturation step is carried out by heating
the reaction mixture to a temperature in the range of 92-95°C; the annealing step
takes place at a temperature in the range of 35-75°C; and the extension step takes
place at a temperature in the range of 65-85°C with a thermal stable DNA
polymerase, such as Taq or Vent (available from Perkin-Elmer Corp., Norwalk, CT,

10   and New England Biolabs, respectively).

Tag complements may be prepared on magnetic microparticles as described
by Albretsen et al, Anal. Biochem., 189: 40-50 (1990), which allows loadings of
several femtomoles of tag complements onto 4.5 µm diameter magnetic beads. Tag
complements may be attached to the microparticles either by their 5' or 3' ends. If

15   attached by 5' ends, then the templates may be sorted via specific hybridization of
tags at their 3' ends. In this embodiment, the template has a primer complement at its
5' end, as shown below:

3'-[oligonucleotide tag]-[template]-[primer complement]-5'

20

The tag complement is then extended the length of the template so that a complement
of the template is obtained which is covalently attached to the microparticle. The
template is removed by heating and the microparticles are washed. After
microparticles are separated, e.g. by flow sorting, repeated cycles of annealing

25   primers, extension, and denaturation are carried out.

If tag complements are attached to the microparticles by their 3' ends, which
allows for convenient synthesis directly on the microparticles, the order of the
oligonucleotide tag and primer complement are reversed, as shown below:

30   5'-[oligonucleotide tag]-[template]-[primer complement]-3'

Also, the 5' end of the tag complement is phosphorylated, e.g. using commercially
available reagents. After specific hybridization via the oligonucleotide tag, a primer
is annealed to the primer complement at the 3' end of the template and extended with

35   a DNA polymerase lacking 3'→5' exonuclease activity. The nick left by this
extension reaction is then ligated and the original template removed by heating. After
separating microparticles, the cycle sequencing can be carried out as above.

Separation of loaded microparticles may be carried out by flow sorting, wherein suspended microparticles are entrained to pass single file through a nozzle and in a liquid jet which is broken up into a regular series of charged droplets which are directed to predetermined target vessels, wells, or other reaction locations on a

5   substrate. Microparticles are conveniently detected in the jet by light scatter and the magnitude of the scatter is used to determine whether a droplet contains no, one, or multiple microparticles. A particularly useful apparatus for such flow sorting and delivery of sequencing reagent is disclosed in Brennan, International application PCT/US94/05896. Once the individual loaded microparticles are distributed to a

10   plurality of reaction sites or wells with the appropriate sequencing reagents, the collection of reactions can be thermally cycled together to generate extension products. After cycling is completed, the extension products are separated by electrophoresis. Preferably, electrophoretic separation is carried out by capillary electrophoresis in a gel-free separation medium, which allows convenient loading and

15   rapid separation of the extension fragments. Also, apparatus is available which permits detection by four-color fluorescence of a large number of samples substantially at the same time, e.g. the type disclosed by Mathies and Huang, Nature, 359:167-169 (1992); Huang et al, Anal. Chem., 64: 2149-2154 (1992); Huang et al, Anal. Chem., 64: 967-972 (1992); or the like. Preferably, several thousand cycle

20   sequencing reactions are carried at the same time. More preferably, mixtures of templates are sorted onto a population of microparticles having a repertoire of oligonucleotide tags of between 1000 and 10,000 different types.

25

Example 1

Synthesis of Oligonucleotide Clamp having 3' and 5'

Cholesterol Binding Moieties

30   The series of oligonucleotide clamps listed in Table I were synthesized that have cholesterol moieties attached to either a 5' end, a 3' end, or to both a 3' end and a 5' end. The 3' cholesterol was attached by first constructing a cholesterol-derivatized solid phase support followed by routine oligonucleotide chain extension via phosphoramidite monomers on a conventional automated DNA synthesizer (Applied

35   Biosystems model 394). The 5' cholesterol was attached in the final coupling step of the synthesis by reacting cholesterol chloroformate with the terminal nucleotide having a 5' amino group or by coupling a cholesterol phosphoramidite with a terminal hydroxyl group, the former method usually giving higher yields. Such clamp are

readily attached to other polymeric units by introducing a free amine group, e.g. via AminoModifier II (Clontech) during synthesis, followed by bromoacetylation, and coupling via a phosphorothioate group, as discussed above.

5   (1) A polymer supported oligonucleotide, 1 μmole scale, with terminal 5'-amino group was treated with 2 ml of a 10% solution of cholesteryl formate in chloroform/diisopropylethylamine (9:1, v:v) for 20 minutes at room temperature. The polymer support was then washed with chloroform and acetonitrile, cleaved and deprotected with concentrated ammonium (5 hours at 55°C), and purified by reverse phase HPLC.

10   (2) A polymer supported oligonucleotide, 1 μmole scale, with terminal hydroxyl group was treated with 250 μl of 0.1 M solution of cholesterol phosphoramidite in chloroform and 250 μl of 0.45 M solution of tetrazole in acetonitrile for 10-15 minutes at room temperature. The polymer support was then washed with acetonitrile, cleaved and deprotected with concentrated ammonium (5

15 hours at 55°C), and purified by reverse phase HPLC.

     The oligonucleotides and oligonucleotide clamps listed in Table I were designed to specifically bind to the following single stranded or double stranded target polynucleotides:

20
```
        5'---AAAAGAAAAGGGGGGA---3'
        3'---TTTTCTTTTCCCCCCT---5'
               Double stranded DNA


        5'---AAAAGAAAAGGGGGGA---3'
```
25   Single stranded RNA or DNA

TABLE V

Designation	Sequence of Oligonucleotide Clamp*
050A	5'-CACTTTTC^{Me}TTTTCCCCCCTCACACTCCCCCCCTTTTCTTTTAC-Chol
050B	5'-Chol-CACTTTTC^{Me}TTTTCCCCCCTCACACTCCCCCCCTTTTCTTTTAC-Chol
051	5'-Chol-CACTTTTC^{Me}TTTTC^{Me}C^{Me}C^{Me}C^{Me}C^{Me}C^{Me}TCACACT-CCCCCCCTTTTCTTTTAC-Chol
052A	5'-CACTTTTC^{Me}TTTTCCCCCCTCACACTC^{Me}C^{Me}C^{Me}C^{Me}C^{Me}C^{Me}TT-TTCTTTTAC-Chol
052B	5'-Chol-CACTTTTC^{Me}TTTTCCCCCCTCACACTC^{Me}C^{Me}C^{Me}C^{Me}C^{Me}C^{Me}TT-TTCTTTTAC-Chol
053	5'-Chol-CACTTTTC^{Me}TTTTGGGGGGTCACACTCCCCCCCTTTTCTTTTAC-Chol
DL015	5'-CACTTTTCTTTTGGGGGGTCACACTCCCCCCCTTTTCTTTTAC
DL014	5'-CACTTTTCTTTTGGGGGGTCACACTCCCCCCCTTTTCTTTTAC-Chol
DL013	5'-Chol-CACTTTTCTTTTGGGGGGTCACACTCCCCCCCTTTTCTTTTAC-Chol
DL021	5'-Chol-TTTTCTTTTCACACTTTTCTTTTGGGGGGTCACACTCCCCCC-Chol
DL022	5'-Chol-CACTTTTCTTTTCCCCCCTCACACACTCCCCCCCTTTTCTTTTAC-Chol
DL023	5'-Chol-TTTTCTTTTCACACTTTTCTTTTCCCCCCTCACTCCCCCC-Chol

* "Chol" represents cholesterol and C^{Me} represents 5-methylated cytidine.

The melting temperature of the following compounds were determined by computing the half maximum of the 260 nm absorption v. temperature curve, as discribed above: DL015: 39.0°C; DL014: 58°C; DL013: 68.0°C; DL021: 67.5°C; DL022: 67.5°C; and control (two unconnected oligonucleotides without binding moities having sequences: 3'-TTTTCTTTTCCCCCCT-5' and 5'-TTTTC^{Me}TTTT(C^{Me})$_6$-3'): 32.0°C.

Example 2

Synthesis of Oligonucleotide Clamp with Cholesterol Binding Moieties, a Polyethylene Glycol Hinge Region, and a Free Amine

5   The following oligonucleotide clamp having a non-nucleosidic hinge region is synthesized as described above. In the hinge region, protected polyethyl glycol phosphoramidites disclosed by Durand et al, Nucleic Acids Research, 18: 6353-6359 (1990); and Rumney and Kool, Angew. Chem. Int. Ed Engl., 31: 1617 (1992) are employed along with AminoModifier II from Clontech (Palo Alto, CA):

10

$5'\text{-Chol-CACTTTTCTTTTGGGGGGp}(OCH_2CH_2)_2pCH_2CH(CH_2NH_2)p\text{-}$

$(OCH_2CH_2)_2pTCCCCCCTTTTCTTTTCA\text{-Chol-}3'$

15   wherein "p" indicates the presence of a phosphodiester linkage.

Example 3

Synthesis of Oligonucleotide Clamp Carrying a Oligonucleotide Attached to Hinge Region

20   The oligonucleotide clamp of Example 2 is synthesized with the free primary amine in its hinge region. The amine is derivatized with bromoacetyl as described above. Separately, an oligonucleotide is prepared having either a 3' or 5' monophosphorothioate group, as desired. The oligonucleotide may remain attached to its synthesis support. The bromoacetylated clamp and the oligonucleotide are
25   combined in an aqueous solution and frozen as described above.

Example 4

Sorting Multiple Target Polynucleotides Derived from pUC19 with Tags having Oligomeric Subunits

30   A mixture of three target polynucleotide-tag conjugates are obtained as follows: First, the following six oligonucleotides are synthesized and combined pairwise to form tag 1, tag 2, and tag 3:

$5'\text{-pTCGACC}(w_1)(w_2)(w_3)(w_4)(w_5)(w_6)(w_7)(w_8)(w_1)A$
35          $GG(**)(**)(**)(**)(**)(**)(**)(**)(**)TTCGAp\text{-}5'$

Tag 1

$$5'-pTCGACC(w_6)(w_7)(w_8)(w_1)(w_2)(w_6)(w_4)(w_2)(w_1)A$$
$$GG(**)(**)(**)(**)(**)(**)(**)(**)(**)TTCGAp-5'$$

Tag 2

$$5'-pTCGACC(w_3)(w_2)(w_1)(w_1)(w_5)(w_8)(w_8)(w_4)(w_4)A$$
$$GG(**)(**)(**)(**)(**)(**)(**)(**)(**)TTCGAp-5'$$

Tag 3

where "p" indicates a monophosphate, the w_i's represent the subunits define in Table I, and the terms "(**)" represent their respective complements. A pUC19 is digested with Sal I and Hind III, the large fragment is purified, and separately ligated with tags 1, 2, and 3, to form pUC19-1, pUC19-2, and pUC19-3, respectively. The three recombinants are separately amplified and isolated, after which pUC19-1 is digested with Hind III and Aat I, pUC19-2 is digested with Hind III and Ssp I, and pUC19-3 is digested with Hind III and Xmn I. The small fragments are isolated using conventional protocols to give three double stranded fragments about 250, 375, and 575 basepairs in length, respectively, and each having a recessed 3' strand adjacent to the tag and a blunt or 3' protruding strand at the opposite end. Approximately 12 nmoles of each fragment are mixed with 5 units T4 DNA polymerase in the manufacturer's recommended reaction buffer containing 33 μM deoxycytosine triphosphate. The reaction mixture is allowed to incubate at 37°C for 30 minutes, after which the reaction is stopped by placing on ice. The fragments are then purified by conventional means.

CPG microparticles (37-74 mm particle size, 500 angstrom pore size, Pierce Chemical) are derivatized with the linker disclosed by Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). After separating into three aliquots, the complements of tags 1, 2, and 3 are synthesized on the microparticles using a conventional automated DNA synthesizer, e.g. a model 392 DNA synthesizer (Applied Biosystems, Foster City, CA). Approximately 1 mg of each of the differently derivatized microparticles are placed in separate vessels.

The T4 DNA polymerase-treated fragments excised from pUC19-1, -2, and -3 are resuspended in 50 μL of the manufacturer's recommended buffer for Taq DNA ligase (New England Biolabs). The mixture is then equally divided among the three vessels containing the 1 mg each of derivatized CPG microparticles. 5 units of Taq DNA ligase is added to each vessel, after which they are incubated at 55°C for 15 minutes. The reaction is stopped by placing on ice and the microparticles are washed

several times by repeated centrifugation and resuspension in TE. Finally, the
microparticles are resuspended in Nde I reaction buffer (New England Biolabs) where
the attached polynucleotides are digested. After separation from the microparticles
the polynucleotide fragments released by Nde I digestion are fluorescently labeled by
5   incubating with Sequenase DNA polymerase and fluorescein labeled thymidine
triphosphate (Applied Biosystems, Foster City, CA). The fragments are then
separately analyzed on a nondenaturing polyacrylamide gel using an Applied
Biosystems model 373 DNA sequencer.

10   **Example 5**
                    Sorting Multiple Target Polynucleotides Derived from pUC19
                          with Tag Complements Consisting of Oligonucleotide
                                    3'N→5'P Phosphoramidates
                    A mixture of three target polynucleotide-tag conjugates are obtained as
15   follows: First, the following six oligonucleotides are synthesized and combined
pairwise to form tag 1, tag 2, and tag 3:

                    5'-pTCGACCGATTGATTTGATTGATATGTA
20   GGCTAACTAAACTAACTATACATTCGAp-5'

                                        Tag 1

                    5'-pTCGACCGTAAAAAGTAGATTTGAGTAA
25   GGCATTTTTCATCTAAACTCATTCGAp-5'

                                        Tag 2

                    5'-pTCGACCTGATTGATATGTATGTGTAAA
30   GGACTAACTATACATACACATTTCGAp-5'

                                        Tag 3

                    Three 20-mer oligonucleotide 3'N→5'P phosphoramidate tag complements are
35   seperately synthesized on CPG microparticles as described above, the tag
complements having the following sequences (in 3'→5' orientation):
CTAACTAAACTAACTATACA, CATTTTTCATCTAAACTCAT, and

ACTAACTATACATACACATT. The polynucleotide-tag conjugates are prepared
and sorted as described in Example 4.

Example 6

Parallel Sequencing of SV40 Fragments

5   A repertoire of 36-mer tags consisting of nine 4-nucleotide subunits selected
from Table I is prepared by separately synthesizing tags and tag complements by a
split and mix approach, as described above. The repertoire is synthesized so as to
permit ligation into a Sma I/Hind III digested M13mp19. Thus, as in Example I, one

10   set of oligonucleotides begins with the addition of A followed by nine rounds of split
and mix synthesis wherein the oligonucleotide is extended subunit-wise by 3'-
phosphoramidite derivatived 4-mers corresponding to the subunits of Table I. The
synthesis is then completed with the nucleotide-by-nucleotide addition of one half of
the Sma I recognition site (GGG), two C's, and a 5'-monophosphate, e.g. via the

15   Phosphate-ON reagent available from Clontech Laboratories (Palo Alto, CA). The
other set of oligonucleotides begins with the addition of three C's (portion of the Sma
I recognition site) and two G's, followed by nine rounds of split and mix synthesis
wherein the oligonucleotide is extended by 3'-phosphoramidite derivatized 4-mers
corresponding to the complements of the subunits of Table I. Synthesis is completed

20   by the nucleotide-by-nucleotide addition of the Hind III recognition site and a 5'-
monophosphate. After separation from the synthesis supports the oligonucleotides
are mixed under conditions that permit formation of the following duplexes:

25   5'-pGGGCC(w$_i$)(w$_i$)(w$_i$)(w$_i$)(w$_i$)(w$_i$)(w$_i$)(w$_i$)(w$_i$)A
       CCCGG(**)(**)(**)(**)(**)(**)(**)(**)(**)TTCGAp-5'

The mixture of duplexes is then ligated into a Sma I/Hind III-digested M13mp19. A

30   repertoire of tag complements are synthesized on CPG microparticles as described
above.

       Next the following adaptor is prepared which contains a Fok I site and
portions of Eco RI and Sma I sites:

35   5'-pAATTCGGATGATGCATGCATCGACCC
                    GCCTACTACGTACGTAGCTGGGp-5'

              Eco RI   Fok I   Sma I

The adaptor is ligated into the Eco RI/Sma I digested M13 described above.

Separately, SV40 DNA is fragmented by sonication following the protocol set forth in Sambrook et al (cited above). The resulting fragments are repaired using standard protocols and separated by size. Fragments in the range of 300-500

5 basepairs are selected and ligated into the Sma I digested M13 described above to form a library of fragment-tag conjugates, which is then amplified. A sample containing several thousand different fragment-tag conjugates is taken from the library, further amplified, and the fragment-tag inserts are excised by digesting with Eco RI and Hind III. The excised fragment-tag conjugates are treated with T4 DNA

10 polymerase in the presence of deoxycytidine triphosphate, as described in Example I, to expose the oligonucleotide tags for specific hybridization to the CPG microparticles.

After hybridization and ligation, as described in Example I, the loaded microparticles are treated with Fok I to produce a 4-nucleotide protruding strand of a

15 predetermined sequence. A 10:1 mixture (probe 1:probe 2) of the following probes are ligated to the polynucleotides on microparticles.

```
        Probe 1         FAM- ATCGGATGAC
                             TAGCCTACTGAGCT

20
        Probe 2         biotin- ATCGGATGAC
                                TAGCCTACTGAGCT
```

FAM represents a fluorescein dye attached to the 5'-hydroxyl of the top strand of

25 Probe 1 through an aminophosphate linker available from Applied Biosystems (Aminolinker). The biotin may also be attached through an Aminolinker moiety and optionally may be further extended via polyethylene oxide linkers, e.g. Jaschke et al (cited above).

The loaded microparticles are then deposited on the surface of an avidinated

30 glass slide to which and from which reagents and wash solutions can be delivered and removed. The avidinated slide with the attached microparticles is examined with a scanning fluorescent microscope (e.g. Zeiss Axioskop equipped with a Newport Model PM500-C motion controller, a Spectra-Physics Model 2020 argon ion laser producing a 488 nm excitation beam, and a 520 nm long-pass emission filter, or like

35 apparatus). The excitation beam and fluorescent emissions are delivered and collected, respectively, through the same objective lens. The excitation beam and collected fluorescence are separated by a dichroic mirror which directs the collected fluorescence through a series of bandpass filters and to photon-counting devices corresponding to the fluorophors being monitored, e.g. comprising Hamamatsu model

9403-02 photomultipliers, a Stanford Research Systems model SR445 amplifier and model SR430 multichannel scaler, and digital computer, e.g. a 486-based computer. The computer generates a two dimensional map of the slide which registers the positions of the microparticles.

5   After cleavage with Fok I to remove the initial probe, the polynucleotides on the attached microparticles undergo 20 cycles of probe ligation, washing, detection, cleavage, and washing, in accordance with the preferred single base sequencing methodology described below. Within each detection step, the scanning system records the fluorescent emission corresponding the base identified at each

10   microparticle. Reactions and washes below are generally carried out with manufacturer's (New England Biolabs') recommended buffers for the enzymes employed, unless otherwise indicated. Standard buffers are also described in Sambrook et al (cited above).

The following four sets of mixed probes are provided for addition to the target

15   polynucleotides:

```
          TAMRA- ATCGGATGACATCAAC
                 TAGCCTACTGTAGTTGANNN
```

20
```
          FAM- ATCGGATGACATCAAC
               TAGCCTACTGTAGTTGCNNN
```

```
          ROX- ATCGGATGACATCAAC
               TAGCCTACTGTAGTTGGNNN
```

25

```
          JOE- ATCGGATGACATCAAC
               TAGCCTACTGTAGTTGTNNN
```

30   where TAMRA, FAM, ROX, and JOE are spectrally resolvable fluorescent labels attached by way of Aminolinker II (all being available from Applied Biosystems, Inc., Foster City, California); the bold faced nucleotides are the recognition site for Fok I endonuclease, and "N" represents any one of the four nucleotides, A, C, G, T. TAMRA (tetramethylrhodamine), FAM (fluorescein), ROX (rhodamine X), and JOE

35   (2',7'-dimethoxy-4',5'-dichlorofluorescein) and their attachment to oligonucleotides is also described in Fung et al, U.S. patent 4,855,225.

The above probes are incubated in approximately 5 molar excess of the target polynucleotide ends as follows: the probes are incubated for 60 minutes at 16°C with 200 units of T4 DNA ligase and the anchored target polynucleotide in T4 DNA ligase

40   buffer; after washing, the target polynucleotide is then incubated with 100 units T4 polynucleotide kinase in the manufacturer's recommended buffer for 30 minutes at

37°C, washed, and again incubated for 30 minutes at 16°C with 200 units of T4 DNA ligase and the anchored target polynucleotide in T4 DNA ligase buffer. Washing is accomplished by successively flowing volumes of wash buffer over the slide, e.g. TE, disclosed in Sambrook et al (cited above). After the cycle of ligation-

5 phosphorylation-ligation and a final washing, the attached microparticles are scanned for the presence of fluorescent label, the positions and characteristics of which are recorded by the scanning system. The labeled target polynucleotide, i.e. the ligated complex, is then incubated with 10 units of Fok I in the manufacturer's recommended buffer for 30 minutes at 37°C, followed by washing in TE. As a result the target

10 polynucleotide is shortened by one nucleotide on each strand and is ready for the next cycle of ligation and cleavage. The process is continued until twenty nucleotides are identified.

APPENDIX I

Exemplary computer program for generating minimally cross hybridizing sets

```
Program minxh
c
c
c
        integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
        dimension nbase(6)
c
c
        write(*,*)'ENTER SUBUNIT LENGTH'
        read(*,100)nsub
100     format(i1)
        open(1,file='sub4.dat',form='formatted',status='new')
c
c
          nset=0
          do 7000 m1=1,3
              do 7000 m2=1,3
                  do 7000 m3=1,3
                      do 7000 m4=1,3
                          sub1(1)=m1
                          sub1(2)=m2
                          sub1(3)=m3
                          sub1(4)=m4
c
c
          ndiff=3
c
c
c                         Generate set of subunits differing from
c                         sub1 by at least ndiff nucleotides.
c                         Save in mset1.
c
c
          jj=1
          do 900 j=1,nsub
900           mset1(1,j)=sub1(j)
c
c
          do 1000 k1=1,3
              do 1000 k2=1,3
                  do 1000 k3=1,3
                      do 1000 k4=1,3
c
c
                          nbase(1)=k1
                          nbase(2)=k2
                          nbase(3)=k3
                          nbase(4)=k4
c
                      n=0
```

```
                         do 1200 j=1,nsub
                           if(sub1(j).eq.1 .and. nbase(j).ne.1 .or.
         1                    sub1(j).eq.2 .and. nbase(j).ne.2 .or.
         3                    sub1(j).eq.3 .and. nbase(j).ne.3) then
                           n=n+1
                           endif
1200                       continue
c
c
         if(n.ge.ndiff) then
c
c
c                                          If number of mismatches
c                                          is greater than or equal
c                                          to ndiff then record
c                                          subunit in matrix mset
c
c
             jj=jj+1
              do 1100 i=1,nsub
1100              mset1(jj,i)=nbase(i)
              endif
c
c
1000     continue
c
c
             do 1325 j2=1,nsub
             mset2(1,j2)=mset1(1,j2)
1325         mset2(2,j2)=mset1(2,j2)
c
c
c                                          Compare subunit 2 from
c                                          mset1 with each successive
c                                          subunit in mset1, i.e. 3,
c                                          4,5, ... etc. Save those
c                                          with mismatches .ge. ndiff
c                                          in matrix mset2 starting at
c                                          position 2.
c                                            Next transfer contents
c                                          of mset2 into mset1 and
c                                          start
c                                          comparisons again this time
c                                          starting with subunit 3.
c                                          Continue until all subunits
c                                          undergo the comparisons.
c
c
         npass=0
c
c
1700     continue
         kk=npass+2
         npass=npass+1
c
c
             do 1500 m=npass+2,jj
```

```
                        n=0
                        do 1600 j=1,nsub
                           if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
          2                    mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
          2                    mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
                              n=n+1
                              endif
      1600                    continue
                        if(n.ge.ndiff) then
                              kk=kk+1
                              do 1625 i=1,nsub
      1625                       mset2(kk,i)=mset1(m,i)
                        endif
      1500              continue
      c
      c
      c                                            kk is the number of subunits
      c                                            stored in mset2
      c
      c
      c
      c                                            Transfer contents of mset2
      c                                            into mset1 for next pass.
      c
      c
      c
                        do 2000 k=1,kk
                           do 2000 m=1,nsub
      2000                    mset1(k,m)=mset2(k,m)
                        if(kk.lt.jj) then
                           jj=kk
                           goto 1700
                           endif
      c
      c
                         nset=nset+1
                        write(1,7009)
      7009               format(/)
                        do 7008 k=1,kk
      7008                 write(1,7010)(mset1(k,m),m=1,nsub)
      7010              format(4i1)
                        write(*,*)
                        write(*,120) kk,nset
      120               format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
      7000               continue
                         close(1)
      c
      c
                        end
```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Sydney Brenner

(ii) TITLE OF INVENTION: Massively parallel sequencing of
sorted polynucleotides

(iii) NUMBER OF SEQUENCES: 17

(iv) CORRESPONDENCE ADDRESS:
        (A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics,
Inc.
        (B) STREET: 3832 Bay Center Place
        (C) CITY: Hayward
        (D) STATE: California
        (E) COUNTRY: USA
        (F) ZIP: 94545

(v) COMPUTER READABLE FORM:
        (A) MEDIUM TYPE: 3.5 inch diskette
        (B) COMPUTER: IBM compatible
        (C) OPERATING SYSTEM: Windows 3.1/DOS 5.0
        (D) SOFTWARE: Microsoft Word for Windows, vers. 2.0

(vi) CURRENT APPLICATION DATA:
        (A) APPLICATION NUMBER:
        (B) FILING DATE:
        (C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
        (A) APPLICATION NUMBER: 08/322,348
        (B) FILING DATE: 13-OCT-94

(vii) PRIOR APPLICATION DATA:
        (A) APPLICATION NUMBER: 08/359,295
        (B) FILING DATE: 19-DEC-94

(viii) ATTORNEY/AGENT INFORMATION:
        (A) NAME: Stephen C. Macevicz
        (B) REGISTRATION NUMBER: 30,285
        (C) REFERENCE/DOCKET NUMBER: mps2wo

(ix) TELECOMMUNICATION INFORMATION:
        (A) TELEPHONE: (510) 670-9365
        (B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 38 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA   38

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 26 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

AATTCGGATG ATGCATGCAT CGACCC   26

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 14 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

TAGCCTACTG AGCT   14

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 16 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: double
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

ATCGGATGAC ATCAAC   16

(2) INFORMATION FOR SEQ ID NO: 5: .

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 16 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

AAAAGAAAAG GGGGGA   16

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 42 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

CACTTTTCTT TTCCCCCCTC ACACTCCCCC CTTTTCTTTT AC   42

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 42 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

CACTTTTCTT TTGGGGGGTC ACACTCCCCC CTTTTCTTTT AC   42

(2) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 42 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

TTTTCTTTTC ACACTTTTCT TTTGGGGGGT CACACTCCCC CC   42

(2) INFORMATION FOR SEQ ID NO: 9:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 14 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

TTTTTTTTTT TTCAC   14

(2) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 14 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

ACTTTTTTTTTTT TTTT   14

(2) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 16 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

CACACACACA CACACACAAC ACACACACAC ACACAC   16

(2) INFORMATION FOR SEQ ID NO: 12:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 14 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear


(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

CAAAAAAAAA AAAC   14

(2) INFORMATION FOR SEQ ID NO: 13:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 17 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

GGACCAATTT CCGGATTCTC GGGAGAG   17

(2) INFORMATION FOR SEQ ID NO: 14:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 16 nucleotides
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

CACATAGGCA TCTCCTAGTG CAGGAAGA   18

(2) INFORMATION FOR SEQ ID NO: 15:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 30 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

TAGGCATCTC CTAGTGCAGG AAGACACACA   30

(2) INFORMATION FOR SEQ ID NO: 16:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 30 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

CACACAAAAA AAAAAAAAAA AAAAAAAAAA   30

(2) INFORMATION FOR SEQ ID NO: 17:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 11 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: double
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

ACCAGCTGAT C   11

I claim:

1.   A method for determining the nucleotide sequence of a target polynucleotide, the method comprising the steps of:

5   generating from the target polynucleotide a plurality of fragments that cover the target polynucleotide;

attaching an oligonucleotide tag from a repertoire of tags to each fragment of the plurality such that substantially all the same fragments have the same oligonucleotide tag attached and substantially all different fragments have different

10   oligonucleotide tags attached, the oligonucleotide tags being selected from the group consisting of oligonucleotides of natural nucleotide monomers having a length in the range of from 10 to 20 nucleotides and oligonucleotides of natural nucleotide monomers comprising a plurality of subunits such that each subunit of the plurality consists of an oligonucleotide of natural nucleotide monomers having a length from

15   three to six nucleotides, the subunits being selected from a minimally cross-hybridizing set;

sorting the fragments by specifically hybridizing the oligonucleotide tags with their respective tag complements;

determining the nucleotide sequence of a portion of each of the fragments of

20   the plurality; and

determining the nucleotide sequence of the target polynucleotide by collating the sequences of the fragments.

2.   The method of claim 1 wherein said complements of said oligonucleotide tags

25   are attached to a solid phase support.

3.   The method of claim 2 wherein said oligonucleotide tags and said complements are single stranded oligonucleotides.

30   4.   The method of claim 3 wherein said step of generating produces randomly overlapping fragments of said target polynucleotide.

5.   The method of claim 4 wherein said step of determining said nucleotide sequence of said fragments are carried out simultaneously for said plurality of

35   fragments by a single base sequencing method.

6.   The method of claim 5 wherein said portions of each of said fragments includes from 12 to 50 nucleotides.

7.   The method of claim 6 wherein said portion of each of said fragments includes from 12 to 25 nucleotides.

5   8.   The method of claim 7 wherein said target polynucleotide is between one and fifty kilobases in length.

9.   The method of claim 5 wherein said solid phase support is a plurality of microparticles each having attached thereto a uniform population of said tag

10   complements.

10.   The method of claim 9 wherein after said step of sorting said plurality of microparticles are fixed to a planar substrate.

15   11.   The method of claim 10 wherein said plurality of microparticles are disposed randomly on the surface of said planar substrate at a density of between about 1000 microparticles to about 100 thousand microparticles per square centimeter.

12.   The method of claim 1 wherein said tag complements of said oligonucleotide

20   tags comprise oligomers of subunits selected from the group consisting of peptide nucleic acids and nucleoside 3'N→5'P phosphoramidates.

13.   The method of claim 12 wherein said complements of said oligonucleotide tags are attached to a solid phase support.

25

14.   The method of claim 13 wherein said oligonucleotide tags and said complements are single stranded oligonucleotides.

15.   The method of claim 14 wherein said step of generating produces randomly

30   overlapping fragments of said target polynucleotide.

16.   The method of claim 15 wherein said step of determining said nucleotide sequence of said fragments are carried out simultaneously for said plurality of fragments by a single base sequencing method.

35

17.   The method of claim 16 wherein said portions of each of said fragments includes from 12 to 50 nucleotides.

18.   The method of claim 17 wherein said portion of each of said fragments includes from 12 to 25 nucleotides.

19.   The method of claim 18 wherein said target polynucleotide is between one and fifty kilobases in length.

20.   The method of claim 19 wherein said solid phase support is a plurality of microparticles each having attached thereto a uniform population of said tag complements.

21.   The method of claim 20 wherein after said step of sorting said plurality of microparticles are fixed to a planar substrate.

22.   The method of claim 21 wherein said plurality of microparticles are disposed randomly on the surface of said planar substrate at a density of between about 1000 microparticles to about 100 thousand microparticles per square centimeter.

23.   The method of claim 1 wherein said complements of said oligonucleotide tags comprise oligonucleotide clamps having at least one oligonucleotide moiety capable of specifically binding to said oligonucleotide tag having at least one pair of binding moieties, each pair of binding moieties comprising a first binding moiety and a second binding moiety, the first binding moiety being covalently linked to an oligonucleotide moiety and the second binding moiety being covalently linked to an oligonucleotide moiety, wherein a stable covalent or non-covalent linkage is formed between the first binding moiety and the second binding moiety of a pair whenever the first and second binding moieties of the pair are brought into juxtapositon by the specific binding to said oligonucleotide tag of the at least one oligonucleotide moiety to which the first and second binding moieties are attached.

24.   The method of claim 23 wherein said oligonucleotide clamp is defined by the formula:

$$X-O_1-G-O_2-Y$$
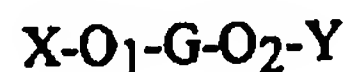
wherein:

O_1 and O_2 are oligonucleotides specific for said oligonucleotide tag;
G is a hinge region; and

X and Y are binding moieties such that X and Y form a stable covalent or non-covalent linkage or complex whenever they are brought into juxtapositon by the annealing of O_1 and O_2 to said oligonucleotide tag.

5   25.   The oligonucleotide clamp of claim 24 wherein:

O_1 and O_2 have the same length and are from about 12 to about 40 nucleotides in length: and

G is defined by the formula:

10 $(M-L)_n$

wherein:

n is between 1 and 10;

L is a linking group; and

15   M is a straight chain, cyclic, or branched organic molecular structure containing from 1 to 20 carbon atoms and from 0 to 10 heteroatoms selected from the group consisting of oxygen, nitrogen, and sulfur.

26.   The oligonucleotide clamp of claim 25 wherein:

20   M is alkyl, alkoxy, alkenyl, or aryl containing from 1 to 16 carbon atoms; heterocyclic having from 3 to 8 carbon atoms and from 1 to 3 heteroatoms selected from the group consisting of oxygen, nitrogen, and sulfur; glycosyl; or nucleosidyl; and

L is a phosphorus(V) linking group.

25

27.   The oligonucleotide clamp of claim 26 wherein:

M is nucleosidyl, or alkyl, alkoxy, or alkenyl containing from 1 to 8 carbon atoms;

L is selected from the group consisting of phosphodiester, phosphotriester,

30   methyl phosphonate, phosphorothioate, and phosphoramidate; and

n is from 4 to 8.

28.   The oligonucleotide clamp of claim 27 wherein X and Y form a hydrophobic complex.

35

29.   The oligonucletide clamp of claim 28 wherein X and Y are each cholesterol.

30.   The method of claim 29 wherein said oligonucleotide clamps are attached to one or more solid phase supports.

31.   The method of claim 30 wherein said step of generating produces randomly
5   overlapping fragments of said target polynucleotide.

32.   The method of claim 31 wherein said step of determining said nucleotide sequence of said fragments are carried out simultaneously for said plurality of fragments by a single base sequencing method.
10

33.   The method of claim 32 wherein said portions of each of said fragments includes from 12 to 50 nucleotides.

34.   The method of claim 33 wherein said target polynucleotide is between one
15   kilobase and fifty kilobases in length.

35.   The method of claim 30 wherein said one or more solid phase supports are a plurality of microparticles each having attached thereto a uniform population of said oligonucleotide clamps and wherein after said step of sorting said plurality of
20   microparticles are fixed to a planar substrate.

36.   A method of classifying a population of polynucleotides, the method comprising the steps of:
        attaching an oligonucleotide tag to each polynucleotide of the population such
25   that substantially all of the same same polynucleotides have the same oligonucleotide tag attached and substantially all different polynucleotides have different oligonucleotide tags attached, the oligonucleotide tags being selected from the group consisting of oligonucleotides of natural nucleotide monomers having a length in the range of from 10 to 20 nucleotides and oligonucleotides of natural nucleotide
30   monomers comprising a plurality of subunits such that each subunit of the plurality consists of an oligonucleotide of natural nucleotide monomers having a length from three to six nucleotides, the subunits being selected from a minimally cross-hybridizing set;
        sorting the polynucleotides by specifically hybridizing the oligonucleotide
35   tags with their respective tag complements;
        determining the nucleotide sequence of a portion of each of the sorted polynucleotides; and

classifying the population of polynucleotides by the frequency distribution of the portions of sequences of the cDNA molecules.

37.   The method of claim 36 wherein said tag complements of said oligonucleotide
5   tags are covalently attached to a solid phase support.

38.   The method of claim 37 wherein said solid phase support is a microparticle and wherein a uniform population of said tag complements is attached to each said microparticle.
10
39.   The method of claim 38 wherein said population of polynucleotides is a cDNA library.

40.   The method of claim 39 wherein said portion of said polynucleotides is in the
15   range of from 12 to 50 nucleotides.

41.   The method of claim 40 wherein said portion of said cDNA molecules is in the range of from 12 to 25 nucleotides.

20   42.   The method of claim 41 wherein said tag complements of said oligonucleotide tags comprise oligomers of subunits selected from the group consisting of peptide nucleic acids and nucleoside 3'N→5'P phosphoramidates.

43.   A method of detecting novel cDNA molecules in a cDNA library, the method
25   comprising the steps of:
        forming a cDNA library from a population of mRNA molecules, each cDNA molecule in the cDNA library having an oligonucleotide tag attached such that substantially all of the same same cDNA molecules have the same oligonucleotide tag attached and substantially all different cDNA molecules have different
30   oligonucleotide tags attached, the oligonucleotide tags being selected from the group consisting of oligonucleotides of natural nucleotide monomers having a length in the range of from 10 to 20 nucleotides and oligonucleotides of natural nucleotide monomers comprising a plurality of subunits such that each subunit of the plurality consists of an oligonucleotide of natural nucleotide monomers having a length from
35   three to six nucleotides, the subunits being selected from a minimally cross-hybridizing set;
        sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective tag complements;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules; and

identifying novel cDNA molecules by comparing the nucleotide sequences of the portions of the sorted cDNA molecules with the sequences of known cDNA

5   molecules.

44.   The method of claim 43 wherein said complements of said oligonucleotide tags are covalently attached to a solid phase support.

10   45.   The method of claim 44 wherein said solid phase support is a microparticle and wherein a uniform population of said complements is attached to each said microparticle.

46.   The method of claim 45 wherein said portion of said cDNA molecules is in
15   the range of from 12 to 50 nucleotides.

47.   The method of claim 46 wherein said portion of said cDNA molecules is in the range of from 12 to 25 nucleotides.

20   48.   The method of claim 47 wherein said tag complements of said oligonucleotide tags comprise oligomers of subunits selected from the group consisting of peptide nucleic acids and nucleoside 3'N→5'P phosphoramidates.

49.   A method of preparing templates for DNA sequencing, the method comprising
25   the steps of:

providing a solution containing a mixture of templates, each template of the mixture having attached an oligonucleotide tag from a repertoire of tags, the oligonucleotide tags being selected from the group consisting of oligonucleotides of natural nucleotide monomers having a length in the range of from 10 to 20 nucleotides and
30   oligonucleotides of natural nucleotide monomers comprising a plurality of subunits such that each subunit of the plurality consists of an oligonucleotide of natural nucleotide monomers having a length from three to six nucleotides, the subunits being selected from a minimally cross-hybridizing set;
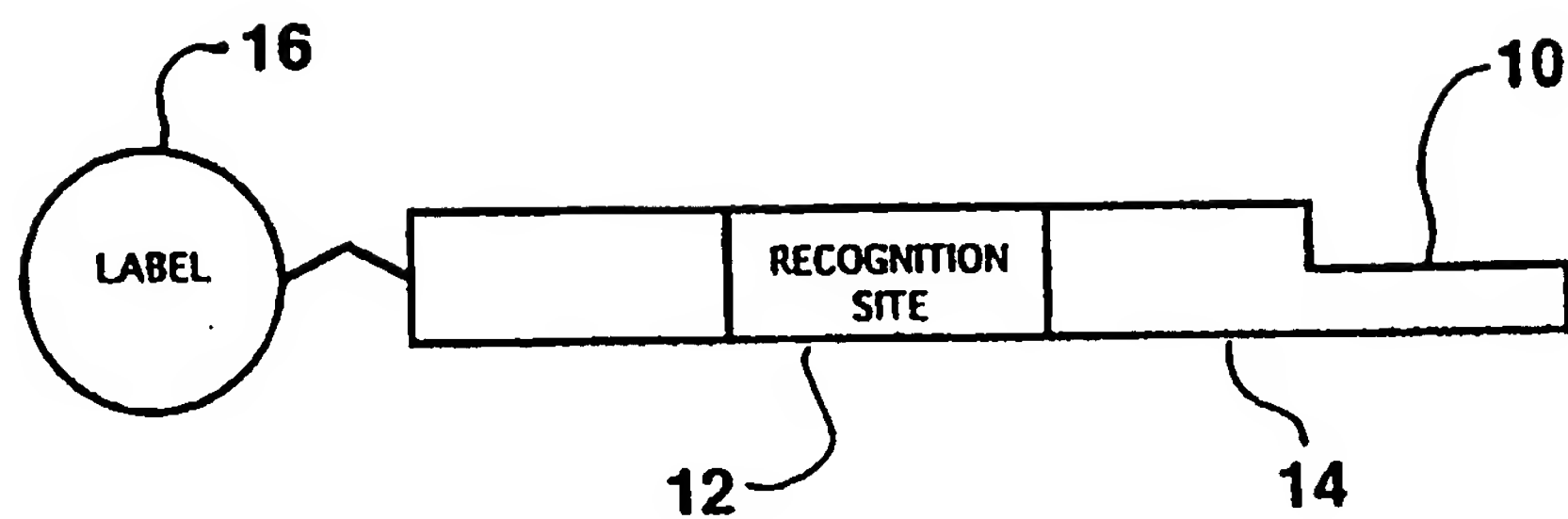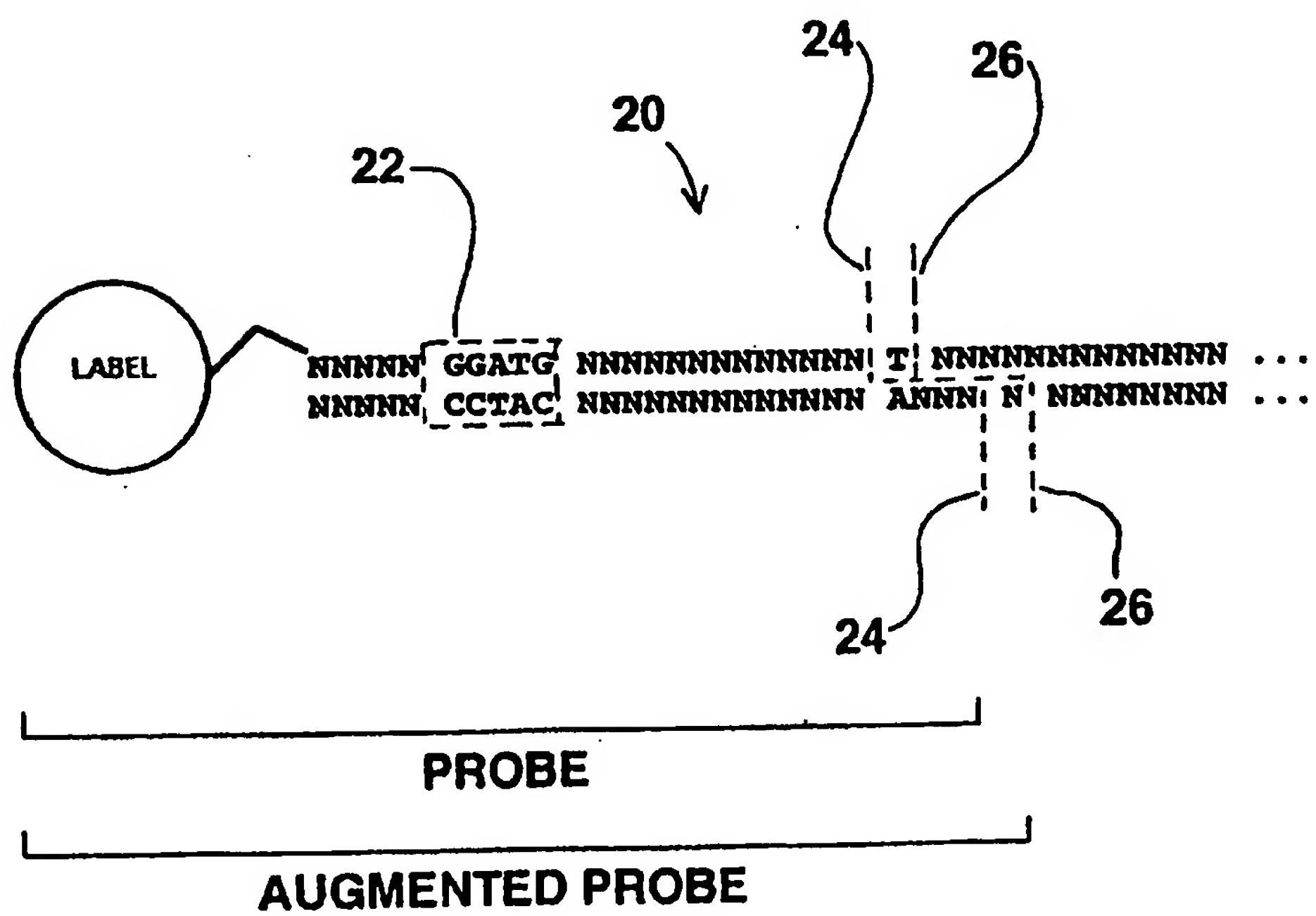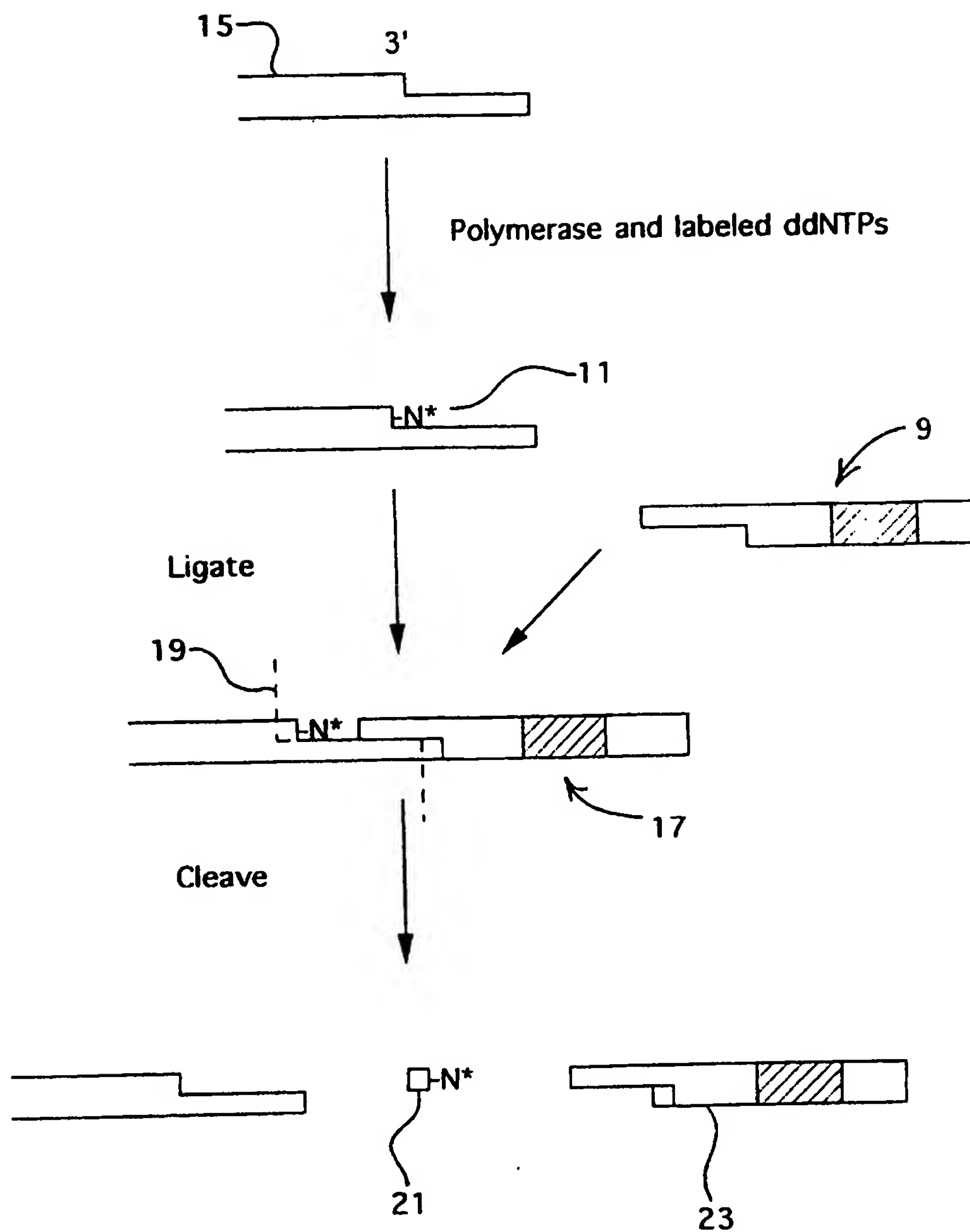
sampling the mixture of templates to form a subpopulation of templates where
35   substantially all templates of the same sequence have the same oligonucleotide tag attached and substantially all templates of different sequences have different oligonucleotide tags attached; and

contacting the subpopulation with one or more solid phase supports having attached thereto tag complements of the oligonucleotide tags under conditions that promote the formation of perfectly matched duplexes between the oligonucleotide tags and their respective complements.

5

50.   The method of claim 49 wherein said one or more solid phase supports are microparticles.

51.   The method of claim 50 wherein said tag complements are selected from the group
10   consisting of peptide nucleic acids and nucleoside 3'N→5'P phosphoramidates whenever said oligonucletide tag is an oligonucleotide of natural nucleotide monomers having a length in the range of from 10 to 20 nucleotides.

52.   The method of claim 50 wherein said oligonucleotide tags comprise a plurality of
15   subunits such that each subunit of the plurality consists of an oligonucleotide of natural nucleotide monomers having a length of from three to six nucleotides, the subunits being selected from a minimally cross-hybridizing set.

Fig. 1a
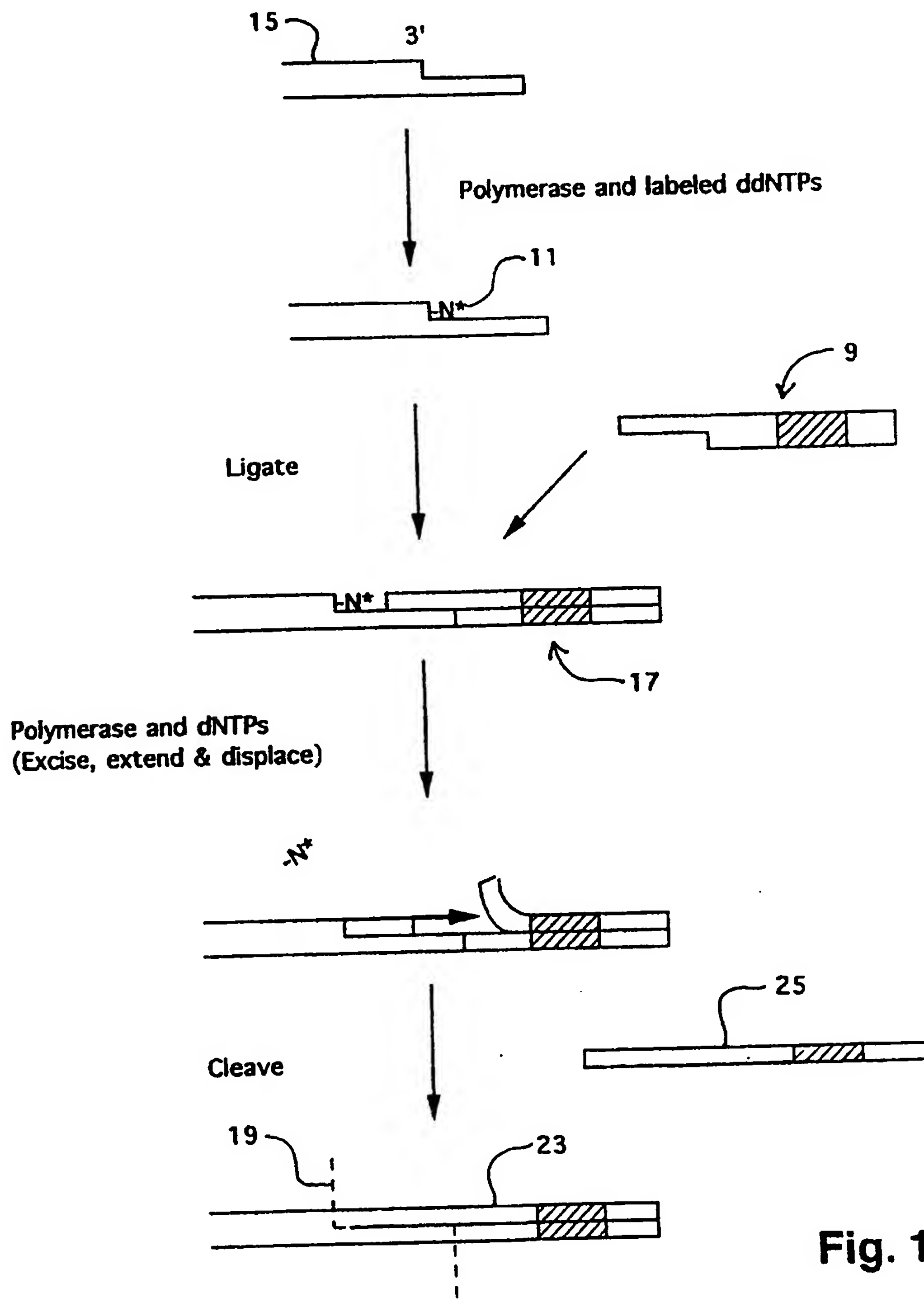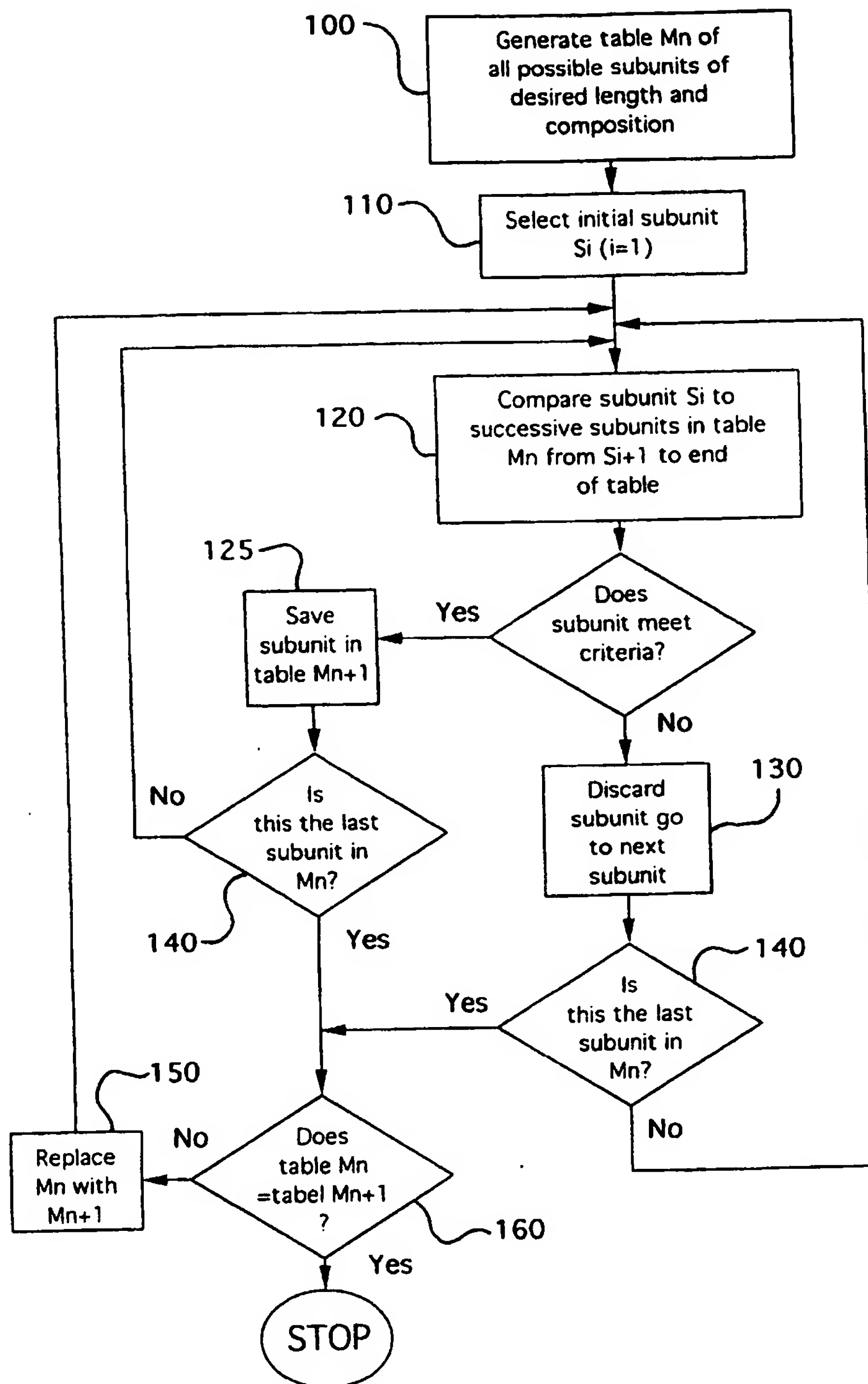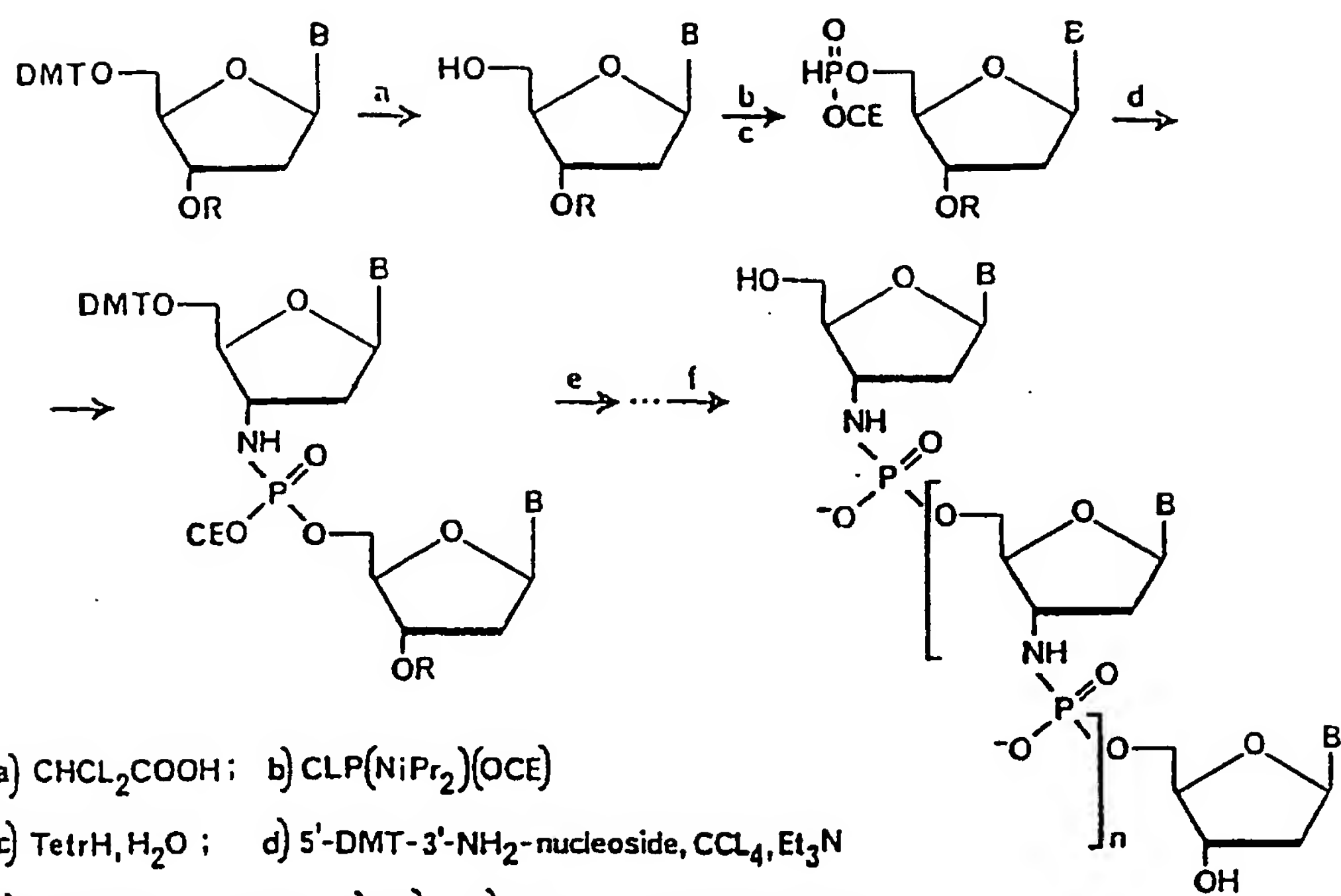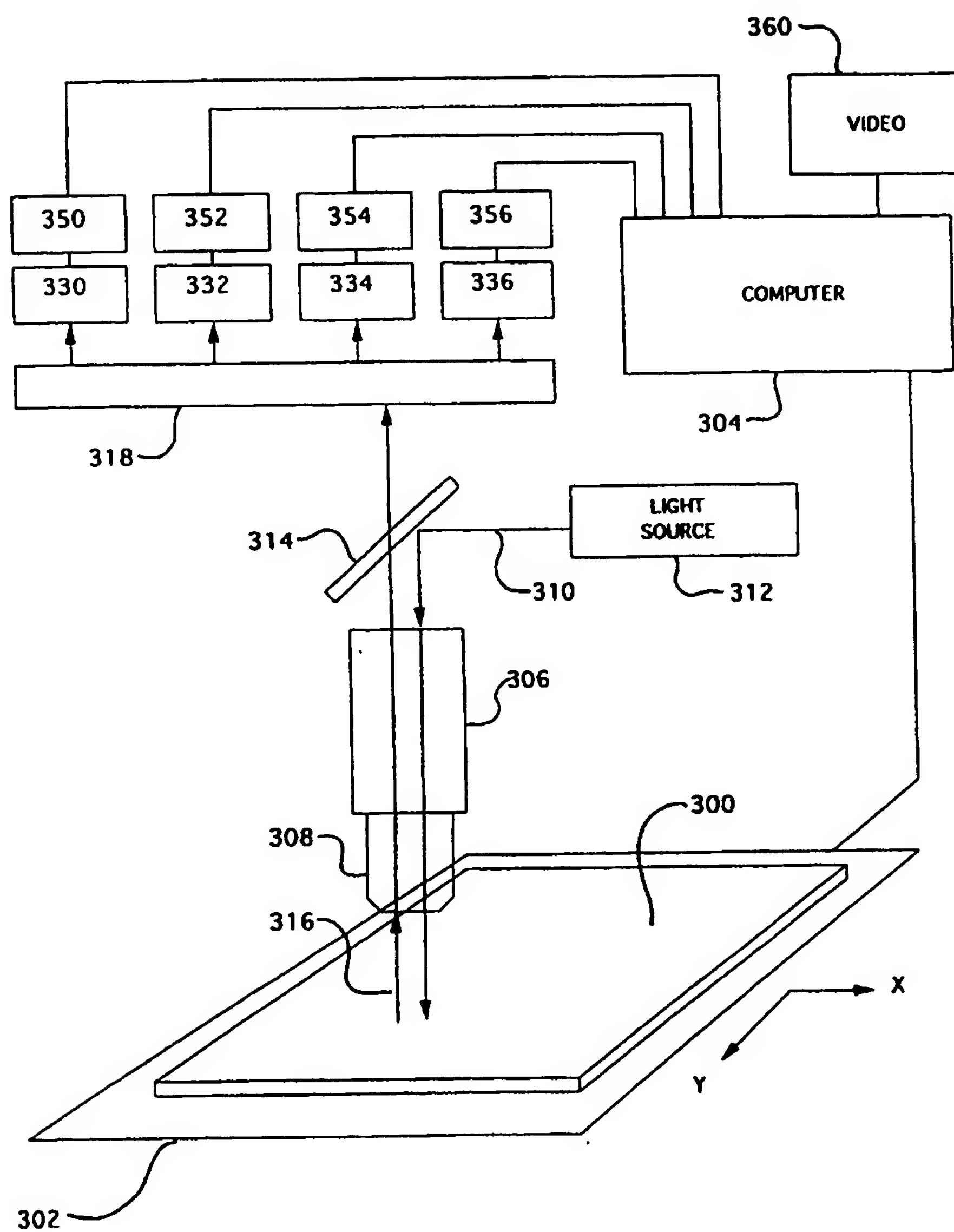
PROBE

AUGMENTED PROBE

Fig. 2

15

3'

↓ Polymerase and labeled ddNTPs

N* — 11

9

Ligate

19

N*

17

Cleave

↓

N*

21   23

Fig. 1b

15 ⌐   3'

↓ Polymerase and labeled ddNTPs

⌐11
⌐N*

9

Ligate ↓

-N*   ///

17

Polymerase and dNTPs
(Excise, extend & displace)

-N*

↓

Cleave

25

19 ⌐   ⌐23

Fig. 1c

Fig. 3

Scheme 1

a) $CHCL_2COOH$;   b) $CLP(NiPr_2)(OCE)$

c) $TetrH, H_2O$;   d) 5'-DMT-3'-NH_2-nucleoside, CCL_4, Et_3N

e) repetition of steps a)–d);   f) NH_3;   R – succinyl CPG

Fig. 4

Fig. 5

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6   C12Q1/68   C07H21/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6   C12Q   C07H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NUCLEIC ACIDS RESEARCH, vol. 17, no. 21, 1989 IRL PRESS LIMITED,OXFORD,ENGLAND, pages 8543-8551, W. RYCHLIK AND E. RHOADS 'A computer program for optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification' cited in the application see the whole document ---	1-52
A	WO,A,92 03461 (IXSYS INC) 5 March 1992 see the whole document ---	1-52
A	WO,A,93 17126 (PUBLIC HEALTH RESEARCH INST OF) 2 September 1993 see the whole document ---	1-52

-/--

[X] Further documents are listed in the continuation of box C.   [X] Patent family members are listed in annex.

° Special categories of cited documents :

'A' document defining the general state of the art which is not considered to be of particular relevance

'E' earlier document but published on or after the international filing date

'L' document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

'O' document referring to an oral disclosure, use, exhibition or other means

'P' document published prior to the international filing date but later than the priority date claimed

'T' later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

'X' document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

'Y' document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

'&' document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 March 1996	26. 03. 96

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Hornig, H

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO,A,93 22684 (AFFYMAX TECH NV ;FODOR STEPHEN P A (US); STRYER LUBERT (US)) 11 November 1993 see the whole document ---	1-52
A	WO,A,93 06121 (AFFYMAX TECH NV) 1 April 1993 see the whole document ---	1-52
P,X	WO,A,95 01456 (LYNX THERAPEUTICS INC ;GRYAZNOV SERGEI M (US); LLOYD DAVID H (US)) 12 January 1995 see the whole document ---	25-29
T	WO,A,95 27080 (LYNX THERAPEUTICS INC) 12 October 1995 see the whole document -----	1-52

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
WO-A-9203461	05-03-92	AU-B-	8505191	17-03-92
		CA-A-	2089362	25-02-92
		EP-A-	0544809	09-06-93
		JP-T-	6503809	28-04-94
		US-A-	5264563	23-11-93
WO-A-9317126	02-09-93	AU-B-	3728093	13-09-93
		CA-A-	2130562	02-09-93
		EP-A-	0675966	11-10-95
WO-A-9322684	11-11-93	AU-B-	4229693	29-11-93
		EP-A-	0639276	22-02-95
WO-A-9306121	01-04-93	AU-B-	2661992	27-04-93
		CA-A-	2118806	01-04-93
		EP-A-	0604552	06-07-94
WO-A-9501456	12-01-95	US-A-	5473060	05-12-95
		AU-B-	7218594	24-01-95
WO-A-9527080	12-10-95	AU-B-	2379195	23-10-95
		CA-A-	2163662	12-10-95